2006 International Symposium on Nonlinear Theory and its Applications

Proceedings

Royal Hotel Carlton, Bologna, Italy
September 11-14, 2006

Organizer:
Research Society of Nonlinear Theory and its Applications, IEICE

In cooperation with:
Technical Group on Nonlinear Problems, IEICE
Technical Group on Circuits and Systems, IEICE
Circuits and Systems Society, IEEE
ARCES, University of Bologna
ENDIF, University of Ferrara
This page has been intentionally left blank.
Greetings & foreword

On behalf of the Organizing Committee, we are extremely glad to welcome you to Bologna and to the 2006 International Symposium on Nonlinear Theory and its Applications (NOLTA2006). This is the 16th Symposium sponsored by the Research Society of Nonlinear Theory and its Applications, IEICE, and, more recently, by the Circuits and Systems Society, IEEE. Since its foundation, NOLTA has been held in Japan, USA, Switzerland, Germany, China and Belgium. We consider a great honour and privilege to organize the 2006 edition in Italy and offer you, for the first time, a taste of the Italian hospitality.

Italy is world-renowned for its history, arts and culture. Cities as Rome, Florence and Venice are everyday the target of thousands of visitors and tourists. Despite not as famous, the beautiful city of Bologna was chosen to host this edition of NOLTA also for its long lasting cultural tradition. Bologna, in fact, possesses the first University in the western world, which was funded in 1088. The history of the city and of its University is rich of great thinkers in science and humanities, making it an indispensable point of reference in the panorama of the European culture. Dante Alighieri, Francesco Petrarcha, Guido Guinizelli, Cino da Pistoia, Cecco d’Ascoli, Re Enzo, Salimbene da Parma and Coluccio Salutati all studied in Bologna, as well as Paracelso, Raimundo de Pegafort, Albrecht Dürer, St. Carlo Borromeo, Torquato Tasso and Carlo Goldoni.

Beside its cultural tradition, Bologna is also famous for its numerous historical buildings and art masterpieces as well as its delicious cuisine. NOLTA2006 will offer to its participants a taste of both during the “Bologna night tour” on September 12th and the conference dinner at “Villa Bassi” on September 13th, that are only a part of the social programme offered by this year’s Symposium.

The most important aspect of NOLTA is however its technical programme. From this point of view, we are very pleased to see that we attracted a very high number of paper submissions. We highly appreciate the invaluable cooperation of three colleagues of ours as Technical Program Committee Co-Chairs, Professors Shinji Doi (Osaka University), Takashi Hisakado (Kyoto University), and Gianluca Mazzini (Ferrara University), who put together a rich in content and high-quality technical program spanning two and a half days of 5 lecture sessions running in parallel and covering all subjects of interest in the field of nonlinear Science and Engineering. Thanks to the efforts of our Special Sessions Co-Chairs, Professors Taishin Nomura (Osaka University) and Riccardo Rovatti (Bologna University), we are also pleased to announce the presence of 25 high-level special sessions covering many cross-disciplinary aspects of nonlinear theory and its applications. Finally, for the first time, the technical programme includes three “Work-in-Progress” sessions: we believe that this is an important addition that aims to give young researchers the opportunity to present the latest results of their work in the qualified NOLTA forum.

A special thank is devoted to authors, special session organizers and plenary speakers for their valuable contributions and to all the Technical Programme Committee members and the reviewers for their professional contributions. We express our gratitude to all the Organizing Committee members, and especially to the three Program Co-Chairs, the two Special Session Co-Chairs and the two Publications Co-Chairs for their dedication and efforts without which the symposium would not have been possible. Thanks also to the Bologna and Ferrara Universities with their army of volunteers: we are very appreciative of their help.

We hope you will have fruitful discussions at the Symposium and that your stay in Bologna will be both rewarding and memorable.

Gianluca Setti and Toshimitsu Ushio
NOLTA2006 General Co-Chairs
On behalf of the Technical Committee of this year’s International Symposium on Nonlinear Theory and its Applications, we are delighted to welcome all of you to Bologna. This year’s NOLTA conference is particularly exciting, both technically and culturally.

The conference includes traditional concentration areas of bifurcation and chaos, analysis and design of nonlinear oscillators and theory and application of signal processing, along with more exotics topics such as complex systems and random number generators. There is much for everyone. The quality of papers is particularly high, and we are confident that the conference attendees will enjoy the technical program. In addition to the regular program, we have 25 exciting special sessions on a variety of topics dealing with the application of nonlinear science to biological- and nano-systems, macroeconomics, complex networks as well as multidimensional information networks.

The very strong program is composed of 265 papers spread over 60 sessions in three and a half long and interesting days. Authors from around the world will be attending and presenting their research results. The papers originate from 30 different Countries, making this symposium the place for learning and interacting with important researchers in the areas of nonlinear theory and its applications. Regular papers have been accepted on the basis of full paper submissions with an acceptance rate of 80% for the review process. The special session contributions were by invitation only. We thank the many special session organizers for their great efforts and initiative in setting up this large variety of interesting sessions. We want to take the opportunity to thank once more everyone who contributed to make this symposium a memorable event: all participants, authors, reviewers, secretaries and organizing committee chairs.

We are also thrilled to mention that we have four outstanding keynote presentations from worldwide experts in the field of nonlinear Science and Engineering: Prof. Steven Boyd (Stanford University), speaking on nonlinear convex optimization, Prof. Yoshimasa Nakamura (Kyoto University), presenting efficient singular value decomposition algorithms, Prof. Sergio Rinaldi (Milano Polytechnic), talking on the goal of evolution in biological network, and Prof. Karl Sigmund (Vienna University), lecturing about the cycles in evolutionary game theory. We are very pleased to have these speakers presenting their ideas and vision at this year’s symposium.

Thank you very much again for coming. Enjoy the technical program, have fruitful discussions, and most of all, have a great time.

Shinji Doi, Takashi Hisakado, and Gianluca Mazzini
NOLTA2006 Technical Programme Co-Chairs
# Table of contents

- **Greetings & foreword** ...................................................... v
- **Table of contents** .......................................................... vii
- **Special thanks** ............................................................... 1
- **Symposium Committees** .................................................. 2
  - Organizing Committee .......................................................... 2
  - Technical Program Committee ............................................... 2
  - Advisory Program Committee ............................................... 3
  - Administrative Committee .................................................... 3
- **Conference venue** .......................................................... 4
- **Conference schedule** ...................................................... 5
- **Sessions at a glance** ...................................................... 6
- **Conference program** ....................................................... 9
  - Day I, September 11th, 2006 .................................................. 9
  - Day II, September 12th, 2006 ............................................... 9
  - Day III, September 13th, 2006 ............................................. 20
  - Day IV, September 14th, 2006 ............................................. 30
- **Conference papers** ........................................................ 37
- **Special Session Organizer index** .................................... 1083
- **Chairman index** ........................................................... 1085
- **Author index** ............................................................... 1087
This page has been intentionally left blank.
Special thanks

The NOLTA2006 Organizing Committee wishes to acknowledge the technical support of the following Institutions which greatly contributed to the success of the event:

<table>
<thead>
<tr>
<th>Technical Group on Nonlinear Problems, IEICE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical Group on Circuits and Systems, IEICE</td>
</tr>
<tr>
<td>Circuits and Systems Society, IEEE</td>
</tr>
<tr>
<td>ARCES, University of Bologna</td>
</tr>
<tr>
<td>ENDIF, University of Ferrara</td>
</tr>
</tbody>
</table>

Finally, we wish to express our deepest thanks and appreciation to the following Institutions, which financially supported the Symposium and greatly contributed to make it a memorable event:

<table>
<thead>
<tr>
<th>Support Center for Advanced Telecommunications Technology Research, Foundation (SCAT), Japan</th>
</tr>
</thead>
<tbody>
<tr>
<td>CARISBO Bank, Italy</td>
</tr>
<tr>
<td>Regione Emilia Romagna, Italy</td>
</tr>
<tr>
<td>University of Bologna, Italy</td>
</tr>
</tbody>
</table>
Symposium Committees

Organizing Committee

GENERAL CO-CHAIRS
Gianluca Setti (University of Ferrara)      Toshimitsu Ushio (Osaka University)

TECHNICAL PROGRAM CO-CHAIRS
Shinji Doi (Osaka University)      Takashi Hisakado (Kyoto University)
Gianluca Mazzini (University of Ferrara)

SPECIAL SESSION CO-CHAIRS
Riccardo Rovatti (University of Bologna)      Taishin Nomura (Osaka University)

PUBLICITY CO-CHAIRS
Mario di Bernardo (University of Naples “Federico II”)      Yoshifumi Nishio (Tokushima University)

PUBLICATION CO-CHAIRS
Sergio Callegari (University of Bologna)      Kaoru Inoue (Doshisha University)

FINANCE CHAIR
Takashi Hikihara (Kyoto University)

LOCAL ARRANGEMENT CHAIR
Stefano Vitali (University of Bologna)

SECRETARIES
Mitsuji Muneyasu (Kansai University)      Masahiro Wada (Konan University)
Tatushi Yamasaki (Kwansei Gakuin University)      Hiroto Tanaka (Osaka University)

STEERING COMMITTEE CHAIR
Kohshi Okumura (Hiroshima Inst. of Tech.) (until May 2006)
Tetsuro Endo (Meiji University) (from June 2006)

HONORARY CHAIR
Kazuo Horiuchi (Waseda University)

Technical Program Committee

CO-CHAIRS
Shinji Doi (Osaka University)      Takashi Hisakado (Kyoto University)
Gianluca Mazzini (University of Ferrara)

MEMBERS
M. Adachi (Tokyo Denki University)      G. Chen (City University of Hong Kong)
S. Doi (Osaka University)      Y. Endo (University of Tsukuba)
O. Feely (University College Dublin)      H. Fujisaka (Hirosima City University)
T. Funaki (Kyoto University)      L. Goldgeisser (Synopsys Inc.)
M. Hasler (EPFL)      Y. Horio (Tokyo Denki University)
K. Ikeda (Kyoto University)      T. Ikeguchi (Saitama University)
Y. Inoue (Waseda University)      K. Jin’no (Kanto Gakuin University)
Y. Kanzawa (Shibaura Institute of Technology)      M. P. Kennedy (University College Cork)
L. Kocarev (University of California San Diego)      K. Konishi (Osaka Prefecture University)
G. M. Maggio (STMicroelectronics)      C. Mocenni (University of Siena)
N. Murata (Waseda University)      H. Nakajima (Kinki University)
K. Natsume (Kyushu Institute of Technology)      Y. Nishio (Tokushima University)
M. Ogorzalek (AGH University of Sci. and Tech.)      Y. Oohama (Kyushu University)
K. Pakdaman (Institut Jacques Monod)      T. Saito (Hosei University)
M. Sano (The University of Tokyo)      S. Sato (Tohoku University)
T. Schimming (OCTAX Microscience GmbH)      J. Suykens (Katholieke Universiteit Leuven)
N. Takahashi (Kyushu University)      H. Tanaka (The University of Electro-Communications)
Y. Tanji (Kagawa University)      K. Terada (Toho University School of Medicine)
I. Tokuda (Japan Advanced Inst. of Sci. and Tech.)      L. Trajkovic (Simon Fraser University)
C. K. Tse (Hong Kong Polytechnic University)      A. Uchida (Takushoku University)
T. Ueta (Tokushima University)      T. Ushio (Osaka University)
H.-F. Yanai (Ibaraki University)      K. Yoshimura (NTT Communication Science Labs)
T. Yoshinaga (Tokushima University)
Advisory Committee

MEMBERS

K. Aihara (University of Tokyo)
G. Chen (City University Hong Kong)
R. Eberhart (IUPUI)
A. Fettweis (Ruhr University Bochum)
W. J. Freeman (U. C. Berkeley)
H. Hayashi (Kyushu Inst. Tech.)
K. Judd (University of Western Australia)
M. P. Kennedy (University College Cork)
E. S. Kuh (U. C. Berkeley)
T. Matsumoto (Fukui University of Tech.)
S. Mori (Keio University)
T. Nagashima (Muroran Inst. of Tech.)
J. A. Nossek (Tech. University of Munich)
S. Oishi (Waseda University)
T. Roska (Hungarian Academy of Science)
T. Saito (Hosei University)
Y. Sawada (Tohoku Inst. of Tech.)
R. Stoop (ETH / University of Zurich)
L. Trajkovic (Simon Fraser University)
Y. Ueda (Future University-Hakodate)
J. Vandewalle (KU Leuven)
A. N. Willson, Jr. (UCLA)
S. Amari (FRP, RIKEN)
L. O. Chua (U. C. Berkeley)
T. Endo (Meiji University)
L. Fortuna (University of Catania)
M. Hasler (EPLF)
M. Iri
H. Kawakami (Tokushima University)
T. Kohda (Kyushu University)
R. W. Liu (University of Notre Dame)
I. Mezic (UCSB)
K. Nakajima (Tohoku University)
T. Nishi (Kyushu University)
M. J. Ogorzalek (AGH University Sci. and Tech.)
M. Plum (Karlsruhe University)
S. M. Rump (Technical University of Humburg)
I. W. Sandberg (University of Texas at Austin)
W. Schwarz (Tech. University of Dresden)
M. Tanaka (Sophia University)
C. K. Tse (Hong Kong Polytech. University)
A. Ushida (Tokushima Bunri University)
P. Werbos (National Science Foundation)
A. H. Zemanian (State University NY at Stony Brook)

Administrative Committee of the Research Society of Nonlinear Theory and its Applications

CHAIR
Tetsuro Endo (Meiji University)

SECRETARY
Yoshifumi Nishio (Tokushima University)

TREASURER
Kenya Jin'no (Kanto Gakuin University)

MEMBERS

K. Aihara (The University of Tokyo)
H. Asai (Shizuoka University)
T. Ushio (Osaka University)
S. Oishi (Waseda University)
K. Okumura (Hiroshima Institute of Tech.)
K. Konishi (Osaka Prefecture University)
T. Saito (Hosei University)
N. Takahashi (Kyushu University)
Y. Tanji (Kagawa University)
K. Nakajima (Tohoku University)
T. Hikihara (Kyoto University)
M. Makino (Chuo University)
K. Yamamura (Chuo University)
Conference venue & floor plan
# Conference schedule

## September 11, 2006 (Monday)

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00-18:30</td>
<td>San Giovanni in Monte cloister</td>
<td>Registrations</td>
</tr>
<tr>
<td>15:00-15:30</td>
<td>Aula Absidale Santa Lucia</td>
<td>Opening ceremony</td>
</tr>
<tr>
<td>15:30-16:30</td>
<td>Aula Absidale Santa Lucia</td>
<td>Plenary talk</td>
</tr>
<tr>
<td>17:00-18:00</td>
<td>Aula Absidale Santa Lucia</td>
<td>Plenary talk</td>
</tr>
<tr>
<td>18:30-21:00</td>
<td>San Giovanni in Monte cloister</td>
<td>Welcome reception</td>
</tr>
</tbody>
</table>

## September 12, 2006 (Tuesday)

**Royal Hotel Carlton**

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00</td>
<td>2nd floor</td>
<td>Registrations</td>
</tr>
<tr>
<td>08:30-10:10</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Morning session I</td>
</tr>
<tr>
<td>10:10-10:30</td>
<td>2nd floor</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:30-12:10</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Morning session II</td>
</tr>
<tr>
<td>12:10-14:00</td>
<td></td>
<td>Lunch break</td>
</tr>
<tr>
<td>14:00-16:00</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Afternoon session I</td>
</tr>
<tr>
<td>16:00-16:20</td>
<td>2nd floor</td>
<td>Coffee break</td>
</tr>
</tbody>
</table>

## September 13, 2006 (Wednesday)

**Royal Hotel Carlton**

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30-10:10</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Morning session I</td>
</tr>
<tr>
<td>10:10-10:30</td>
<td>2nd floor</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:30-12:10</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Morning session II</td>
</tr>
<tr>
<td>12:10-14:00</td>
<td></td>
<td>Lunch break</td>
</tr>
<tr>
<td>14:00-16:00</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Afternoon session I</td>
</tr>
<tr>
<td>16:00-16:20</td>
<td>2nd floor</td>
<td>Coffee break</td>
</tr>
<tr>
<td>19:45</td>
<td></td>
<td>Bus leaving for the Gala Dinner site</td>
</tr>
<tr>
<td>20:30</td>
<td>Villa Bassi</td>
<td>Gala dinner</td>
</tr>
</tbody>
</table>

## September 14, 2006 (Thursday)

**Royal Hotel Carlton**

<table>
<thead>
<tr>
<th>Time</th>
<th>Room</th>
<th>Schedule</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30-10:10</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Morning session I</td>
</tr>
<tr>
<td>10:10-10:30</td>
<td>2nd floor</td>
<td>Coffee break</td>
</tr>
<tr>
<td>10:30-12:10</td>
<td>Imperiale A, Imperiale B, Saturno A, Giove A, Giove B</td>
<td>Morning session II</td>
</tr>
<tr>
<td>12:10-14:00</td>
<td></td>
<td>Lunch break</td>
</tr>
<tr>
<td>14:00-15:00</td>
<td>Imperiale A+B</td>
<td>Plenary talk</td>
</tr>
<tr>
<td>15:30-16:30</td>
<td>Imperiale A+B</td>
<td>Plenary talk</td>
</tr>
<tr>
<td>16:30-16:45</td>
<td>Imperiale A+B</td>
<td>Closing remarks</td>
</tr>
</tbody>
</table>
## Sessions at a glance

### September 11, 2006 (Monday)

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>15:00-15:30</td>
<td>Opening ceremony (Aula absidale Santa Lucia)</td>
</tr>
<tr>
<td>15:30-16:30</td>
<td>Plenary talk (Aula absidale Santa Lucia)</td>
</tr>
<tr>
<td>17:00-18:00</td>
<td>Plenary talk (Aula absidale Santa Lucia)</td>
</tr>
<tr>
<td>18:30-21:00</td>
<td>Welcome reception (San Giovanni in Monte cloister)</td>
</tr>
</tbody>
</table>

### September 12, 2006 (Tuesday)

<table>
<thead>
<tr>
<th>Time</th>
<th>Imperiale A</th>
<th>Imperiale B</th>
<th>Saturno A</th>
<th>Giove A</th>
<th>Giove B</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30</td>
<td>Synchronization I</td>
<td>Special Session SS08</td>
<td>Special Session SS14</td>
<td>Neural Networks I</td>
<td>Complex Systems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hybrid Systems I (Modeling and Behavioral Analysis)</td>
<td>Fuzzy Measures and Cooperative Games</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td>Special Session SS14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td>Neural Networks I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td>Complex Systems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td>Special Session SS14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td>Neural Networks II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Synchronization II</td>
<td>Special Session SS10</td>
<td>Special Session SS05</td>
<td>Special Session SS01</td>
<td>Special Session SS01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hybrid Systems I (Phenomena)</td>
<td>Consistency and Synchronization</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td></td>
<td></td>
<td>Neural Networks II</td>
<td>Special Session SS07</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Special Session SS11</td>
<td>Special Session SS10</td>
<td>Special Session SS05</td>
<td>Nonlinear dynamics in</td>
<td>Nonlinear dynamics in</td>
</tr>
<tr>
<td></td>
<td>Nonlinear Time Series Analysis I</td>
<td>Hybrid Systems III (Design, Control and Applications)</td>
<td>Consistency and Synchronization</td>
<td>Macroeconomics I</td>
<td>Macroeconomics II</td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td></td>
<td>Neural Networks II</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:20</td>
<td></td>
<td></td>
<td>Special Session SS01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:20</td>
<td></td>
<td></td>
<td>Computer Assisted Proof for Linear and Nonlinear Problems in R^n</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:20</td>
<td>Special Session SS12</td>
<td>Random Number Generators</td>
<td>Special Session SS03</td>
<td>Special Session SS02</td>
<td>Special Session SS17</td>
</tr>
<tr>
<td>Time</td>
<td>Imperiale A</td>
<td>Imperiale B</td>
<td>Saturno A</td>
<td>Giove A</td>
<td>Giove B</td>
</tr>
<tr>
<td>--------</td>
<td>-------------------------------------------------</td>
<td>-------------------------------------------------</td>
<td>------------------------------------------------</td>
<td>----------------------------------------------</td>
<td>----------------------------------------------</td>
</tr>
<tr>
<td>08:30</td>
<td><strong>Special Session SS13</strong></td>
<td><strong>Special Session SS23</strong></td>
<td><strong>Special Session SS15</strong></td>
<td>Power Circuits and Systems</td>
<td>Complex Dynamics in Oscillators</td>
</tr>
<tr>
<td></td>
<td>Topological and Computational Approaches to Nonlinear Dynamics and Bifurcations</td>
<td>Multidimensional Mobile Information Networks I</td>
<td>Neurodynamics and its Hardware Implementation I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Coffee break</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td><strong>Special Session SS15</strong></td>
<td><strong>Special Session SS16</strong></td>
<td>Neural Networks III</td>
<td>Analysis of Oscillators</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multidimensional Mobile Information Networks II</td>
<td>Neurodynamics and its Hardware Implementation I</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td><strong>Special Session SS25</strong></td>
<td>Discrete Time Chaos</td>
<td>Special Session SS21</td>
<td>Cellular Neural Networks</td>
<td>Special Session SS19</td>
</tr>
<tr>
<td></td>
<td>Complex Networks Applications</td>
<td></td>
<td>Nonlinear Speech and Image Processing</td>
<td></td>
<td>Rhythmic and Synchronous Activities in Biological Systems</td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:20</td>
<td><strong>Analysis of Dynamical and Uncertain Systems</strong></td>
<td>Chaos Communication</td>
<td>Image Processing</td>
<td>Theory and Applications of Signal Processing</td>
<td>Optimization</td>
</tr>
<tr>
<td>18:20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>September 13, 2006 (Wednesday)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Time</th>
<th>Imperiale A</th>
<th>Imperiale B</th>
<th>Saturno A</th>
<th>Giove A</th>
<th>Giove B</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:30</td>
<td>Control</td>
<td><strong>Special Session SS04</strong></td>
<td>Mathematical Physics</td>
<td>Communications</td>
<td>Work in Progress I</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nonlinear Dynamics in Nanosystems</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Coffee break</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td><strong>Special Session SS18</strong></td>
<td><strong>Special Session SS18</strong></td>
<td>Chaos and Bifurcation</td>
<td>Work in Progress II</td>
<td>Work in Progress III</td>
</tr>
<tr>
<td></td>
<td>Advanced Modeling and Simulation Techniques</td>
<td>Nonlinear Information Processing in the Auditory Feedback Loop</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td></td>
<td></td>
<td></td>
<td>Plenary talk ( Room: Imperiale A + B)</td>
<td></td>
</tr>
<tr>
<td>15:00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15:30</td>
<td></td>
<td></td>
<td></td>
<td>Plenary talk ( Room: Imperiale A + B)</td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Closing remarks ( Room: Imperiale A + B)</td>
</tr>
<tr>
<td>16:45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| September 14, 2006 (Thursday)                                                                                                                                  |
This page has been intentionally left blank.
Conference program

Day I, September 11th 2006

PLENARY SESSION – I

Date: Sep, 11th  
Time: 15:30 – 16:30  
Room: Aula Absidale S. Lucia

Chaired by: Gianluca Setti (University of Ferrara)

15:30 – 16:30  Advances in Nonlinear Convex Optimization  
Stephen P. Boyd (Stanford University)

PLENARY SESSION – II

Date: Sep, 11th  
Time: 17:00 – 18:00  
Room: Aula Absidale S. Lucia

Chaired by: Toshimitsu Ushio (Osaka University)

17:00 – 18:00  New Singular Value Decomposition Algorithm with High Performance in Terms of Integrable Systems  
Yoshimasa Nakamura (Graduate School of Informatics, Kyoto University and PRESTO JST, Japan)

Day II, September 12th 2006

Synchronization – I

Date: Sep, 12th  
Time: 08:30 – 10:10  
Room: Imperiale A

Chaired by: Martin Hasler (EPFL), Masahiro Wada (Konan University)

08:30 – 08:50  Synchronization Regions and Behaviour of Nonlinear Periodic Circuits  
Giancarlo S. Gajani (Polytechnic of Milano), Angelo Brambilla (Polytechnic of Milano)

08:50 – 09:10  Synchronization of Chaotic Circuits Linked by Cross Talk  
Yuki Nakaaji (Tokushima University), Yoshifumi Nishio (Tokushima University)

09:10 – 09:30  Dual Synchronization of Chaos in Mackey-Glass Electronic Circuits with Time Delay  
Satoshi Sano (Takushoku University), Atushi Uchida (Takushoku University), Shigeru Yoshimori (Takushoku University)

09:30 – 09:50  A Master Stability Function Approach for Separation and Synchronization of Chaotic Systems  
Arturo Buscarino (University of Catania), Luigi Fortuna (University of Catania), Mattia Frasca (University of Catania)

09:50 – 10:10  Synchronous Regimes in a Chain of Coupled Bonhoeffer-van der Pol Oscillators  
Grigory Osipov (Nizhny Novgorod University), Alexey Kryukov (Nizhny Novgorod University), Andrey Polovinkin (Nizhny Novgorod University), Jürgen Kurths (University of Potsdam)
**SPECIAL SESSION**

**SS08 — Hybrid Systems – I (Modeling and Behavioral Analysis) ......................... 59**

**Date:** Sep, 12th  **Time:** 08:30 – 10:10  **Room:** Imperiale B

**Organized by:** Kazutoshi Gohara (Hokkaido University), Toshimichi Saito (Hosei University), Wolfgang Schwarz (Dresden University of Technology), Chi K. Tse (Hong Kong Polytechnic University), Toshimitsu Ushio (Osaka University)

**Chaired by:** Wolfgang Schwarz (Dresden University of Technology)

08:30 – 08:50 Modeling and Analysis of a Class of Hybrid Systems by Return Maps .................. 59

   Jörg Krupar (Dresden University of Technology), Axel Schild (Ruhr-Universitaet Bochum), Wolfgang Schwarz (Dresden University of Technology), Jan Lunze (Ruhr-Universitaet Bochum)

08:50 – 09:10 Classification and Analytical Method of Switched Dynamical Systems .................. 63

   Yue Ma (The University of Tokushima), Soumitro Banerjee (Indian Institute of Technology), Takuji Kousaka (Oita University), Tetsushi Ueta (The University of Tokushima)

09:10 – 09:30 Mixed-Signal Modeling and Design Aspects of Clock-Recovery Circuits and Continuous-Time Sigma-Delta Modulators ......................................................... 67

   Florian A. Probst (University of Hannover), Judith Schmackers (University of Hannover), Wolfgang Mathis (University of Hannover)

09:30 – 09:50 Switching Dynamics, Chaos, and Singular Perturbation Analysis of Heart Arrhythmia .............................. 71

   Naoyuki Takahashi (Osaka University), Shinji Doi (Osaka University), Sadatoshi Kumagai (Osaka University)

09:50 – 10:10 Predicting Instability of Power Systems based on Hybrid System Reachability Analysis . 75

   Yoshihiko Susuki (Kyoto University), Takashi Hikihara (Kyoto University), Hiroaki Ebina (Kyoto University)

**SPECIAL SESSION**

**SS14 — Fuzzy Measures and Cooperative Games ................................. 79**

**Date:** Sep, 12th  **Time:** 08:30 – 10:10  **Room:** Saturno A

**Organized by:** Masayo Tsurumi (Osaka University), Masahiro Inuiguchi (Osaka University)

**Chaired by:** Masahiro Inuiguchi (Osaka University), Masayo Tsurumi (Osaka University)

08:30 – 08:50 Fuzzy Measures and Integrals in Re-identification Problems ............................. 79

   Jordi Nin (Artificial Intelligence Research Institute IIIA-CSIC), Vicenç Torra (Artificial Intelligence Research Institute IIIA-CSIC)

08:50 – 09:10 Generalizing Choquet Integral with Level Dependent Capacities ......................... 83

   Salvatore Greco (University of Catania), Silvio Giove (University of Venice), Benedetto Matarazzo (University of Catania)

09:10 – 09:30 Attribute Importance in the Setting of Variable Precision Rough Sets .................. 87

   Masahiro Inuiguchi (Osaka University), Masayo Tsurumi (Osaka University)

09:30 – 09:50 Solutions of Cooperative Fuzzy Games Extended from Ordinary Cooperative Games .............................. 91

   Tetsuzo Tanino (Osaka University)

09:50 – 10:10 A Solution for Fuzzy Generalized Multialternative Games ............................. 95

   Masayo Tsurumi (Osaka University), Masahiro Inuiguchi (Osaka University), Tetsuzo Tanino (Osaka University)
Neural Networks – I

Date: Sep, 12th
Time: 08:30 – 10:10
Room: Giove A

Chaired by: Takahiro Nakanishi (Iwate University), Seiichiro Moro (University of Fukui)

08:30 – 08:50 Property of Random Series from Chaos Neural Network
Takahiro Nakanishi (Iwate University), Hitoaki Yoshida (Iwate University), Satoshi Kawamura (Ishinomaki Senshu University)

08:50 – 09:10 Fractal Analysis of Chaos Neural Network Outputs in Transient State and Steady State
Hitoaki Yoshida (Iwate University), Osamu Ohira (Iwate University), Hiroto Taira (Iwate University), Takahiro Nakanishi (Iwate University)

09:10 – 09:30 Solving Ability of Hopfield Neural Network for QAP by Changing Chaotic Behavior of Switching Noise
Yoshifumi Tada (Tokushima University), Yoko Uwate (Tokushima University), Yoshifumi Nishio (Tokushima University)

09:30 – 09:50 Quinary Multiplier LOGO Neural Networks Based on Mixed Radices
Ali M. Haidar (Beirut Arab University), Hassan A. Osseily (Beirut Arab University), Elias Nassar (Notre Dame University), Hiroyuki Shirahama (Matsuyama University)

09:50 – 10:10 Global Path Searching Based on Local Corporative and Competitive Circuits on Variable Resistive Network
Masatoshi Sato (Sophia University), Hisashi Aomori (Sophia University), Mamoru Tanaka (Sophia University)

Complex Systems

Date: Sep, 12th
Time: 08:30 – 10:10
Room: Giove B

Chaired by: Oscar De Feo (University College Cork), Yasuteru Hosokawa (Shikoku University)

08:30 – 08:50 Identifying Returns Distribution by Using Mixture Distribution Optimized by Genetic Algorithm and Its Applications
Kangrong Tan (Kurume University), Shozo Tokinaga (Kyushu University)

08:50 – 09:10 A Decision Based Model of Addiction: Explaining Quitting Behavior
Chiara Mocenni (University of Siena), Giuseppe Montefrancesco (University of Siena), Silvia Tiezzi (University of Siena)

09:10 – 09:30 Target and Spiral Waves in Cardiac Cultures
Oleg Kanakov (Nizhny Novgorod University), Grigory Osipov (Nizhny Novgorod University), Chi-Keung Chan (Academia Sinica)

09:30 – 09:50 A Basic Theory for Available Operation of Extremely Complicated Large-Scale Network Systems
Kazuo Horiuchi (Waseda University)

09:50 – 10:10 Analysis of Multi-fractal Signals in Artificial Markets Generated by Multi-Agents Using the Genetic Programming for Learning and its Applications
Yoshikazu Ikeda (Shinshu University), Shozo Tokinaga (Kyushu University)
Synchronization – II

Date:  Sep, 12th  Time:  10:30 – 12:10  Room:  Imperiale A
Chaired by:  Takaya Miyano (Ritsumeikan University), Masahiro Nakagawa (Nagaoka University of Technology)

10:30 – 10:50 Robust Stochastic Direct Adaptive Radial Basis Neural Networks Synchronization of Chua's Circuits with Unknown Function ................................. 139
Iman Saboori (The Amirkabir University of Technology), Mohammad B. Menhaj (The Amirkabir University of Technology), Sayyed K. Nikravesh (The Amirkabir University of Technology), Bahram Karimi (The Amirkabir University of Technology)

10:50 – 11:10 Instabilities of Synchronization in Ensembles of Circle Maps ....................... 143
Grigory Osipov (Nizhny Novgorod University), Mikhail Ivanchenko (Nizhny Novgorod University), Jürgen Kurths (University of Potsdam)

11:10 – 11:30 A New Spatio-temporal Chaotic System with Advantageous Synchronization and Unpredictability Features ..................................... 147
Soumaya Meherzi (SYSCOM/ENIT, LSS/SUPELEC/CNRS), Sylvie Marcos (LSS/SUPELEC/CNRS), Safya Belghith (SYSCOM/ENIT)

11:30 – 11:50 Chaotic Synchronization in a Network of Living Mouse Pancreatic B-Cells .............. 151
Takaya Miyano (Ritsumeikan University), Kensuke Ninomiya (Ritsumeikan University), Takahiro Kanno (Fukushima Medical University), Makoto Wakui (Hirosaki University School of Medicine)

11:50 – 12:10 Performance Analysis for Scheme of Synchronization-Manifold Shift Keying ........... 155
Thang M. Hoang (Nagaoka University of Technology), Masahiro Nakagawa (Nagaoka University of Technology)

SPECIAL SESSION

SS09 — Hybrid Systems – II (Phenomena) ........................................ 159

Date:  Sep, 12th  Time:  10:30 – 12:10  Room:  Imperiale B
Organized by:  Kazutoshi Gohara (Hokkaido University), Toshimichi Saito (Hosei University), Wolfgang Schwarz (Dresden University of Technology), Chi K. Tse (Hong Kong Polytechnic University), Toshimitsu Ushio (Osaka University)
Chaired by:  K. Gohara (Hokkaido University), C. K. Tse (Hong Kong Polytechnic University)

10:30 – 10:50 Complex Intermittency in Switching Converters ................................. 159
Yufei Zhou (Anhui University), Herbert H.C. Lu (University of Western Australia), Chi K. Tse (Hong Kong Polytechnic University), Jun-Ning Chen (Anhui University)

10:50 – 11:10 Experimental Confirmation of the Torus Birth through Border-collision Bifurcation in Voltage Controlled PWM-1 Buck Converter ............................. 163

11:10 – 11:30 Nonlinear Phenomena in Hybrid Replicator Dynamics with Changes of Interpretation Functions ....................................................... 167
Takafumi Kanazawa (Osaka University), Toshimitsu Ushio (Osaka University)

11:30 – 11:50 Fractals in Hybrid Systems ........................................ 171
Kazutoshi Gohara (Hokkaido University)

11:50 – 12:10 Interaction of Fast-Scale and Slow-Scale Bifurcations in Current-Mode Controlled DC/DC Converters ................................. 175
Yanteng Chen (Hong Kong Polytechnic University), Chi K. Tse (Hong Kong Polytechnic University), Shui-Sheng Qiu (South China University of Technology)
Analysis and Design of Oscillators .......................................................... 179

Date: Sep, 12th Time: 10:30 – 12:10 Room: Saterno A
Chaired by: Angelo Brambilla (Polytechnic of Milano), Tetsuro Endo (Meiji University)

10:30 – 10:50 Averaging Method Analysis for the Nonlinearly-Coupled van der Pol and Rayleigh Oscillators .......................................................... 179
Toshihiro Hishida (Meiji University), Kuniyasu Shimizu (Meiji University), Tetsuro Endo (Meiji University)

10:50 – 11:10 Averaging Method Analysis for Three-Coupled Van Der Pol Oscillators with Nonlinear Coupling ......................................................... 183
Kuniyasu Shimizu (Meiji University), Tetsuro Endo (Meiji University), Hisa-Aki Tanaka (The University of Electro-Communications)

11:10 – 11:30 Didactic Methods of Synchronization of Continuous Nonlinear Systems .................. 187
Ioan Grosu (University of Medicine and Pharmacy Gr. T. Popa), Anca Lazar (University of Medicine and Pharmacy Gr. T. Popa), Servilia Oancea (University of Agriculture and Veterinary Medicine)

11:30 – 11:50 Transient Analysis for Phase Pattern Switching in Pulse-Driven Star-Coupled Wien-Bridge Oscillators ......................................................... 191
Seichiro Moro (University of Fukui), Tadashi Matsumoto (Fukui University of Technology)

11:50 – 12:10 Nonlinear Analysis of Varactor-tuned Voltage-Controlled Oscillators ....................... 195
Antonio Buonomo (Second University of Naples), Alessandro Lo Schiavo (Second University of Naples), Massimiliano Martucci (STMicroelectronics)

Neural Networks – II .................................................................................. 199

Date: Sep, 12th Time: 10:30 – 12:10 Room: Giove A
Chaired by: Ruedi Stoop (ETHZ), Tadashi Tsubone (Nagaoka University of Technology)

10:30 – 10:50 Shooting SOM and its Application for Clustering ............................................... 199
Masato Tomita (Tokushima University), Haruna Matsushita (Tokushima University), Yoshifumi Nishio (Tokushima University)

10:50 – 11:10 A Novel Logic Oriented Artificial Neural Network to Implement Sequential Logic Circuits 203
Ali M. Haidar (Beirut Arab University)

11:10 – 11:30 Reunifying Self-Organizing Map ...................................................... 207
Haruna Matsushita (Tokushima University), Yoshifumi Nishio (Tokushima University)

11:30 – 11:50 3D Mapping and Ranking Scheme Projection for the Self-Organizing Maps ............ 211
Daisuke Shima (Sophia University), Kaname Kurokawa (Sophia University), Hisashi Aomori (Sophia University), Mamoru Tanaka (Sophia University)

11:50 – 12:10 Chaotically Oscillating Sigmoid Function in Feedforward Neural Network ............... 215
Yoko Uwate (Tokushima University), Yoshifumi Nishio (Tokushima University)
SPECIAL SESSION

SS06 — Nonlinear Dynamics in Macroeconomics – I .............................................. 219

Date: Sep, 12th Time: 10:30 – 12:10 Room: Giove B
Organized by: Akio Matsumoto (Chuo University)
Chaired by: Akio Matsumoto (Chuo University)

10:30 – 10:50 Comparative Dynamics for Climate Policy Choice .............................................. 219
Yasuo Nonaka (Kanto Gakuin University)

10:50 – 11:10 A Simple Structure of Bifurcating Economies .............................................. 223
Junichi Minagawa (Chuo University)

11:10 – 11:30 Price Flexibility and Instability in a Macrodyn(807)mic Model with a Debt Effect .............. N/A
Toichiro Asada (Chuo University), Toshio Inaba (Waseda University), Tetsuya Misawa (Nagoya City University)

11:30 – 11:50 Induced Technical Change and Economic Fluctuations: An Imperfect Competition Model ................................................................. 227
Hiroyuki Yoshida (Nihon University)

11:50 – 12:10 Simple Business Cycle Model with Policy Lags .............................................. 231
Akio Matsumoto (Chuo University), Mami Suzuki (Aichigakusen University)

SPECIAL SESSION

SS11 — Nonlinear Time Series Analysis – I ................................................................. 235

Date: Sep, 12th Time: 14:00 – 16:00 Room: Imperiale A
Organized by: Michael Small (Hong Kong Polytechnic University), Yoshito Hirata (Tokyo University), Tomomichi Nakamura (Hong Kong Polytechnic University)
Chaired by: Michael Small (The Hong Kong Polytechnic University), Tomomichi Nakamura (The Hong Kong Polytechnic University), Yoshito Hirata (The University of Tokyo)

14:00 – 14:20 Detection and Quantification of Directional Coupling Using Recurrences .............................................. 235
Maria C. Romano (University of Potsdam), Marco Thiel (University of Potsdam), Jürgen Kurths (University of Potsdam)

14:20 – 14:40 Direct or Indirect? Graphical Models for Phase Synchronization .............................................. 239
Björn Schelter (University of Freiburg), Matthias Winterhalder (University of Freiburg), Jürgen Kurths (University of Potsdam), Rainer Dahlhaus (University of Heidelberg), Jens Timmer (University of Freiburg)

14:40 – 15:00 Twin Surrogates to Test for Complex Synchronization .............................................. 243
Marco Thiel (University of Potsdam), Maria C. Romano (University of Potsdam), Jürgen Kurths (University of Potsdam)

15:00 – 15:20 Assessing Dependences within Multivariate Time Series Partializing the Knowledge of Thirds ................................................................. 247
Cristian Carmeli (Swiss Federal Institute of Technology of Lausanne), Oscar De Feo (University College Cork)

15:20 – 15:40 Investigating the Limitations of Compartmental Disease Models with Kalman Filters ... 251
David M. Walker (Hong Kong Polytechnic University), Heung W. J. Lee (Hong Kong Polytechnic University), Michael Small (Hong Kong Polytechnic University)

15:40 – 16:00 Deterministic Chaos of Human Cardiac Signals .............................................. 255
Yi Zhao (Hong Kong Polytechnic University), Michael Small (Hong Kong Polytechnic University)
**SPECIAL SESSION**

**SS10 — Hybrid Systems – III (Design, Control, and Applications)** ........................................ 259

**Date:** Sep, 12th **Time:** 14:00 – 16:00 **Room:** Imperiale B

**Organized by:** Kazutoshi Gohara (*Hokkaido University*), Toshimichi Saito (*Hosei University*), Wolfgang Schwarz (*Dresden University of Technology*), Chi K. Tse (*Hong Kong Polytechnic University*), Toshimitsu Ushio (*Osaka University*), Shigemasa Takai (*Kyoto Institute of Technology*), Naly Rakoto-Ravalontsalama (*Ecole des Mines de Nantes*)

**Chaired by:** Toshimitsu Ushio (*Osaka University*), Hiroyuki Torikai (*Hosei University*)

14:00 – 14:20 Design of a Low EMI Hysteretic Current-Controlled DC/DC Boost Converter Via Chaotic Perturbation ......................................................... 259

Michele Balestra (*University of Ferrara*), Fabio Pareschi (*University of Ferrara*), Gianluca Setti (*University of Ferrara*), Riccardo Rovatti (*University of Bologna*)

14:20 – 14:40 A Digital Spiking Neuron and its Simple Learning ............................................. 263

Hiroyuki Torikai (*Hosei University*), Yoshiaki Shimizu (*Tokyo Metropolitan Government*), Toshimichi Saito (*Hosei University*)

14:40 – 15:00 Probabilistic Controllability Analysis of Piecewise Affine Systems .......................... 267

Shun-ichi Azuma (*Kyoto University*), Jun-ichi Imura (*Tokyo Institute of Technology*)

15:00 – 15:20 Controller Optimisation for Hybrid Systems by use of Return Maps ....................... 271

Jörg Krupar (*Dresden University of Technology*), Wolfgang Schwarz (*Dresden University of Technology*)

15:20 – 15:40 Supervisory Control of an Experimental Switched DES ......................................... 275


15:40 – 16:00 Control-invariance of Hybrid Systems with Sampled-data state Feedback Control .... 279

Yoshiyuki Tsuchie (*Osaka University*), Toshimitsu Ushio (*Osaka University*)

**SPECIAL SESSION**

**SS05 — Consistency and Synchronization** ................................................................. 283

**Date:** Sep, 12th **Time:** 14:00 – 16:00 **Room:** Saturno A

**Organized by:** Atsushi Uchida (*Takushoku University*)

**Chaired by:** Kazuyuki Yoshimura (*NTT Corporation*), Atsushi Uchida (*Takushoku University*)

14:00 – 14:20 Consistency of a Laser System Driven by Chaos and Noise Signals ....................... 283

Atsushi Uchida (*Takushoku University*), Shigeru Yoshimori (*Takushoku University*), Ryan McAllister (*University of Maryland*), Rajarshi Roy (*University of Maryland*)

14:20 – 14:40 Consistency in Time-delay Systems with Periodic Feedback Functions .................... 287


14:40 – 15:00 Consistency and ICA of Solvable Chaotic Maps with Non-Gaussian Stable Distributions 291

Ken Umeno (*National Institute of Information and Communications Technology*), Atsushi Uchida (*Takushoku University*), Minghui Kao (*Waseda University*)

15:00 – 15:20 Reproducibility of Limit-cycle Oscillators Induced by Random Impulses .......... 295

Kensuke Arai (*Kyoto University*), Hiroya Nakao (*Kyoto University*)

15:20 – 15:40 Multiple Basins of Consistency in a Driven Mackey-Glass Electronic Circuit ........... 299

Hoipang Yip (*Takushoku University*), Satoshi Sano (*Takushoku University*), Atsushi Uchida (*Takushoku University*), Shigeru Yoshimori (*Takushoku University*)
15:40 – 16:00 Consistency Caused by Various Types of Signals ........................................... 303
Shin-itiro Goto (NTT Communication Science Laboratories), Kazuyuki Yoshimura (NTT Communication Science Laboratories)

SPECIAL SESSION

SS01 — Computer Assisted Proof for Linear and Nonlinear Problems in \(\mathbb{R}^n\) ......... 307

Date: Sep, 12th Time: 14:00 – 16:00 Room: Giove A
Organized by: Shin’ichi Oishi (Waseda University), Siegfried M. Rump (Hamburg University of Technology)
Chaired by: Shin’ichi Oishi (Waseda University), Siegfried M. Rump (Hamburg University of Technology)

14:00 – 14:20 Numerical Verification of Five Solutions in Two-transistor Circuits .................. 307
Yusuke Nakaya (Waseda University), Shin’ichi Oishi (Waseda University), Tetsuo Nishi (Waseda University), Martin Claus (Dresden University of Technology)

14:20 – 14:40 Error Bounds for Extremely Ill-conditioned Problems ................................... 311
Siegfried M. Rump (Hamburg University of Technology)

14:40 – 15:00 Convergence Theorem of Rump’s Method for Inverting Arbitrarily Ill-Conditioned Matrices ................................................................. 315
Shin’ichi Oishi (Waseda University), Kunio Tanabe (Waseda University), Takeshi Ogita (Waseda University), Siegfried M. Rump (Hamburg University of Technology / Waseda University)

15:00 – 15:20 Fast Verification for Sparse Linear Systems with Generalized Diagonally Dominant Matrices ................................................................. 319
Takeshi Ogita (JST / Waseda University), Shin’ichi Oishi (Waseda University)

15:20 – 15:40 Adaptive Verification Method for Dense Linear Systems ................................. 323
Katsuhisa Ozaki (Waseda University), Takeshi Ogita (JST / Waseda University), Shin’ichi Oishi (Waseda University)

15:40 – 16:00 VSDP: A MATLAB Software Package for Verified Semidefinite Programming .... 327
Christian Jansson (Hamburg University of Technology)

SPECIAL SESSION

SS07 — Nonlinear Dynamics in Macroeconomics – II .................................................. 331

Date: Sep, 12th Time: 14:00 – 16:00 Room: Giove B
Organized by: Akio Matsumoto (Chuo University)
Chaired by: Hiroyuki Kurokawa (Kanto-Gakuin University)

14:00 – 14:20 Fractal with Price Scale of High-frequency Data in Futures Market .................. 331
Yoshiaki Kumagai (Waseda University)

14:20 – 14:40 Monetary Policy Subject to Measurement Errors of Private Sector Adaptive Expectations ................................................................. 335
Stefano Casaccia (University of Genova), Silvano Cincotti (University of Genova), Marco Raberto (University of Genova), Andrea Teglio (University of Genova)

14:40 – 15:00 Nonconvex Variational Problem with Recursive Integral Functionals in Sobolev Spaces: Existence and Representation ................................. 339
Nobusumi Sagara (Hosei University)
15:00 – 15:20 Analysis of Artificial Double Auction Markets Consisting of Multi-agents using Genetic Programming for Learning and its Applications to Control
Yoshikazu Ikeda (Shinshu University), Shozo Tokinaga (Kyushu University)

15:20 – 15:40 Fractal Analysis of Euro Foreign Exchange Rates Using the R/S Analysis
Hiroyuki Kurokawa (Kanto Gakuin University)

**SPECIAL SESSION**

**SS12 — Nonlinear Time Series Analysis – II.**

**Date:** Sep, 12th  
**Time:** 16:20 – 18:20  
**Room:** Imperiale A

**Organized by:** Michael Small (Hong Kong Polytechnic University), Yoshito Hirata (Tokyo University), Tomomichi Nakamura (Hong Kong Polytechnic University)

**Chaired by:** Michael Small (The Hong Kong Polytechnic University), Tomomichi Nakamura (The Hong Kong Polytechnic University), Yoshito Hirata (The University of Tokyo)

16:20 – 16:40 Detecting Nonlinearity in Non-stationary Time Series
Tomomichi Nakamura (The Hong Kong Polytechnic University), Michael Small (The Hong Kong Polytechnic University), Yoshito Hirata (The University of Tokyo)

16:40 – 17:00 Inferring on the Dynamical Nature of Spike Trains
Oscar De Feo (University College Cork)

17:00 – 17:20 Phase Space Projection with Time Domain Constraint (TDC) for Noise Reduction
Xiaodong Luo (Hong Kong Polytechnic University), Jie Zhang (Hong Kong Polytechnic University), Michael Small (Hong Kong Polytechnic University)

17:20 – 17:40 Predictive Models of Wind for Controlling a Wind Turbine
Yoshito Hirata (The University of Tokyo), Hideyuki Suzuki (The University of Tokyo), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

17:40 – 18:00 Chaotic Correlation among Cycles in Human Electrocardiograms
Jie Zhang (Hong Kong Polytechnic University), Dong X. Luo (Hong Kong Polytechnic University), Michael Small (Hong Kong Polytechnic University)

**Random Number Generators**

**Date:** Sep, 12th  
**Time:** 16:20 – 18:20  
**Room:** Imperiale B

**Chaired by:** Akio Tsuneda (Kumamoto University), Sergio Callegari (University of Bologna)

16:20 – 16:40 Entropy Enhancement in a Chaos-Based Random Bit Generator
Tommaso Addabbo (University of Siena), Massimo Aiello (University of Siena), Ada Fort (University of Siena), Santina Rocchi (University of Siena), Valerio Vignoli (University of Siena)

16:40 – 17:00 A Study on Generation of Aperiodic Random Numbers Using Tent-Map-Type Feedback Shift Registers
Akio Tsuneda (Kumamoto University), Takashi Tanoue (Kumamoto University), Sho Mitsuishi (Kumamoto University), Daisaburo Yoshioka (Kumamoto University), Takahiro Inoue (Kumamoto University)

17:00 – 17:20 Simple and Effective Post-Processing Stage for Random Stream Generated by a Chaos-Based RNG
Fabio Pareschi (University of Ferrara), Riccardo Rovatti (University of Bologna), Gianluca Setti (University of Ferrara)
17:20 – 17:40 On the Robustness to Noise and Interference of ADC-Derived, Chaos-Based True Random Number Generators.

Sergio Callegari (University of Bologna)

17:40 – 18:00 Compensated True Random Number Generator Based On a Double-Scroll Attractor

Salih Ergun (TUBITAK), Serdar Ozoguz (Istanbul Technical University)

18:00 – 18:20 A Low Cost Chaos-based Random Number Generator Realized in 8-bit Precision Environment

Kwok W. Tang (City University of Hong Kong), Wallace Tang (City University of Hong Kong)

**SPECIAL SESSION**

**SS03 — Discrete Event Systems**

**Date:** Sep, 12th  **Time:** 16:20 – 18:20  **Room:** Saturno A

**Organized by:**  Florian Wenck (Ruhr-Universitaet Bochum), Shigemasa Takai (Kyoto Institute of Technology), Naly Rakoto-Ravalontsalama (Ecole des Mines de Nantes)

**Chaired by:**  Florian Wenck (Ruhr-Universitaet Bochum), Shigemasa Takai (Kyoto Institute of Technology)

16:20 – 16:40 Computational Complexity Analysis of a Petri Net Identification Procedure

Maria P. Cabasino (University of Cagliari), Alessandro Giua (University of Cagliari), Carla Seatzu (University of Cagliari)

16:40 – 17:00 On Reachability Analysis of Multi Agent Nets

Toshiyuki Miyamoto (Osaka University), Masaki Sakamoto (Osaka University), Sadatoshi Kumagai (Osaka University)

17:00 – 17:20 How to Obtain Coefficients for a Firing Count Vector Expanded by T-Invariants and Particular Solutions in P/T Petri Nets

Tadashi Matsumoto (Fukui University of Technology), Seiichiro Moro (University of Fukui), Masahiro Osogami (Fukui University of Technology)

17:20 – 17:40 Computation of Controllable Sublanguages for Unbounded Petri Nets Using Their Approximation Models

Shigemasa Takai (Kyoto Institute of Technology), Yongming Bai (Wakayama University)

17:40 – 18:00 Compositional Controllability Analysis for Partially-Strict Composed DES

Florian Wenck (Ruhr-Universitaet Bochum), Francisco de Assis Carvalho da Silva Neto (Federal University of Santa Catarina)

18:00 – 18:20 A New Algorithm for Diagnosability Analysis of a Class of Hierarchical Finite State Machines

Andrea Paoli (University of Bologna), Stéphane Lafortune (The University of Michigan)
**SPECIAL SESSION**

**SS02 — Computer Assisted Proof for Partial Differential Equations** ........................................ 423

Date:   Sep, 12th  Time:   16:20 – 18:20  Room:  Giove A

Organized by:  Michael Plum (Karlsruhe University), Nobito Yamamoto (The University of Electro-Communications)

Chaired by:   Michael Plum (Karlsruhe University), Nobito Yamamoto (The University of Electro-Communications)

16:20 – 16:40  A New Mixed-type Embedding Theorem ........................................ N/A
    Michael Plum (Universität Karlsruhe)

16:40 – 17:00  A Numerical Verification Method for ODEs based on Nakao’s Theory  ............... 423
    Nobito Yamamoto (The University of Electro-Communications)

17:00 – 17:20  Some Computer Assisted Proofs on Three Dimensional Heat Convection Problems .... 427
    Myoungnyoun Kim (Kyushu University), Mitsuhiro T. Nakao (Kyushu University), Yoshitaka Watanabe (Kyushu University), Takaaki Nishida (Waseda University)

17:20 – 17:40  A Computer-assisted Proof for the Kolmogorov Flows of Incompressible Viscous Fluid   431
    Yoshitaka Watanabe (Kyushu University)

17:40 – 18:00  Enclosures for the Biharmonic Equation ........................................ N/A
    Borbála Fazekas (Universität Karlsruhe)

18:00 – 18:20  Enclosures for Nonlinear Hyperbolic Problems ........................................ N/A
    Vu Hoang (Universität Karlsruhe)

**SPECIAL SESSION**

**SS17 — Laser Dynamics and Nonlinear Optics** ........................................ 435

Date:   Sep, 12th  Time:   16:20 – 18:20  Room:  Giove B

Organized by:  Atsushi Uchida (Takushoku University)

Chaired by:   Atsushi Uchida (Takushoku University)

16:20 – 16:40  Interplay of Optical Feedback and Polarization Dynamics in VCSELs ................. 435
    Krassimir Panajotov (Vrije Universiteit Brussel), Hugo Thienpont (Vrije Universiteit Brussel), Marc Sciamanna (SUPELEC Campus de Metz)

16:40 – 17:00  Bifurcation to Polarization Chaos in VCSELs with Orthogonal Optical Injection ........ 439
    Marc Sciamanna (SUPELEC Campus de Metz), Ignace Gatere (SUPELEC Campus de Metz), Hugo Thienpont (Vrije Universiteit Brussel), Krassimir Panayotov (Vrije Universiteit Brussel)

17:00 – 17:20  Experimental Observation of In-phase and Anti-phase Synchronization of Chaos in Coupled Vertical-cavity Surface-emitting Lasers ........................................ 443
    Mitsutoshi Ozaki (Takushoku University), Takaya Mihara (Takushoku University), Hiroyuki Someya (Takushoku University), Atsushi Uchida (Takushoku University), Shigeru Yoshimori (Takushoku University)

17:20 – 17:40  Chaos Synchronization in Modulated Lasers ........................................ 447
    Riccardo Meucci (CNR-INOA), Francesco Salvadori (University of Florence), Kais Al-Naimee (University of Baghdad), Fortunato T. Arecchi (University of Florence)

17:40 – 18:00  Chaotic Dynamics and Synchronization in Microchip Lasers Coupled with Optoelectronic Feedback ........................................ 451
    Keisuke Mizumura (Takushoku University), Atsushi Uchida (Takushoku University), Shigeru Yoshimori (Takushoku University)
**Day III, September 13th 2006**

**SPECIAL SESSION**

**SS13 — Topological and Computational Approaches to Nonlinear Dynamics and Bifurcations**

- **Date:** Sep, 13th  
- **Time:** 08:30 – 10:10  
- **Room:** Imperiale A

**Organized by:** Hiroshi Kokubu (Kyoto University), Takashi Hikihara (Kyoto University)

**Chaired by:** Hiroshi Kokubu (Kyoto University)

- **08:30 – 08:50** Recent Developments in Topological-computational Approaches to Nonlinear Dynamics: An Overview  
  Hiroshi Kokubu (Kyoto University)

- **08:50 – 09:10** Symbolic Dynamics and Heteroclinic Transitions for the Rossler System  
  Piotr Zgliczynski (Jagiellonian University)

- **09:10 – 09:30** Computing Arnold Tongue Scenarios: Some Recent Advances  
  Frank Schilder (University of Bristol)

- **09:30 – 09:50** An Algorithm for Hyperbolicity Verification and its Application  
  Zin Arai (Kyoto University)

- **09:50 – 10:10** Rigorous Bounds for the Probability of Chaos in the Logistic Family  
  Stefano Luzzatto (Imperial College London)

**SPECIAL SESSION**

**SS23 — Multidimensional Mobile Information Networks – I**

- **Date:** Sep, 13th  
- **Time:** 08:30 – 10:10  
- **Room:** Imperiale B

**Organized by:** Shoji Shinoda (Chuo University)

**Chaired by:** Kyung Sup Kwak (Inha University)

- **08:30 – 08:50** UWB Radio: A Real Chance for Application of Chaotic Communications  
  Geza Kolumban (Budapest University of Technology and Economics), Tamas Krebesz (Budapest University of Technology and Economics)

- **08:50 – 09:10** Ultrawideband Transceiver Platform Based On Chaotic Signals  
  Alexander Dmitriev (Institute of Radio engineering and Electronics of RAS), Andrey Kletsov (Institute of Radio engineering and Electronics of RAS), Lev Kuzmin (Institute of Radio engineering and Electronics of RAS), Anton Laktushkin (Institute of Radio engineering and Electronics of RAS), Anrey Panas (Institute of Radio engineering and Electronics of RAS), Vladimir Sinyakin (Bauman Moscow State Technical University)

- **09:10 – 09:30** On Waiting Time for Path Reconstruction in Mobile Multi-hop Networks  
  Keisuke Nakano (Niigata University), Masakazu Sengoku (Niigata University), Shoji Shinoda (Chuo University)

- **09:30 – 09:50** Distributed Opportunistic Wireless Spectrum Access Protocols  
  Jianhua He (University of Bristol), Alistair Munro (University of Bristol), Dritan Kaleshi (University of Bristol), Joe McGeehan (University of Bristol)

- **09:50 – 10:10** Investigations on Throughput Gain of MIMO Multiplexing Schemes in HSDPA for High-Speed Mobile Multimedia Services  
  Hidekazu Taoka (NTT DoCoMo, Inc.), Hiroyuki Atarashi (NTT DoCoMo, Inc.), Kenichi Higuchi (NTT DoCoMo, Inc.), Mamoru Sawahashi (Musashi Institute of Technology)
SS15 — Neurodynamics and its Hardware Implementation — I  

Date: Sep, 13th  
Time: 08:30 – 10:10  
Room: Saturno A

Organized by: Koji Nakajima (Tohoku University)  
Chaired by: Koji Nakajima (Tohoku University), Shigeo Sato (Tohoku University)

08:30 – 08:50 A CMOS Synapse with STDP Function and Its Application to Hopfield-type Associative Memory  
Hideki Tanaka (Kyushu Institute of Technology), Takashi Morie (Kyushu Institute of Technology), Kazuyuki Aihara (The University of Tokyo/ERATO JST)  

08:50 – 09:10 Physical Chaotic Neuro-Dynamics and Optimization  
Yoshihiko Horio (Tokyo Denki University), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

09:10 – 09:30 Dynamic Behavior and Characteristics of the Modified ID Model with Burst Firing  
Shinya Suenaga (Tohoku University), Yoshihiro Hayakawa (Tohoku University), Koji Nakajima (Tohoku University)

09:30 – 09:50 An STDP-type Learning by Minimizing K-L Divergence for a Spiking Neural Network  
Shigeo Sato (Tohoku University), Kun Ma (Tohoku University), Koji Nakajima (Tohoku University)

Power Circuits and Systems

Date: Sep, 13th  
Time: 08:30 – 10:10  
Room: Giove A

Chaired by: Wolfgang Schwarz (Dresden University of Technology), Kaoru Inoue (Doshisha University)

08:30 – 08:50 Non-linear Model of a Parallel Hybrid Power Audio Amplifier  
Lorenzo Chiesi (University of Parma), Emilio Lorenzani (University of Parma), Giovanni Franceschini (University of Parma), Alberto Bellini (DISMI-University of Modena)

08:50 – 09:10 Design-Oriented Hopf Bifurcation Boundary in Parallel-Connected Buck Converters Under Democratic Current-Sharing Control  
Herbert H.C. Iu (University of Western Australia), Chi K. Tse (Hong Kong Polytechnic University)

09:10 – 09:30 On the Application of a Server Type RTK-GPS to Electric Power Systems  
Masashi Sakamoto (Kagoshima University), Hitoshi Takata (Kagoshima University)

09:30 – 09:50 Effects of Characteristics of Band-pass Filter to Operation of Class DE Inverter  
Motoki Katayama (Chiba University), Hiroyuki Hase (Chiba University), Hiroo Sekiya (Chiba University), Jianming Lu (Chiba University), Takashi Yahagi (Chiba University)

09:50 – 10:10 A Novel Control Method of a Regenerative Power Storage System for Energy Saving of an Electric Machinery  
Kaoru Inoue (Doshisha University), Kenji Ogata (Doshisha University), Toshiji Kato (Doshisha University)
Complex Dynamics in Oscillators ................................................................. 531

Date: Sep, 13th  Time: 08:30 – 10:10   Room: Giove B
Chaired by: Keiji Konishi (Osaka Prefecture University), Hisato Fujisaka (Hiroshima City University)

08:30 – 08:50 Pattern Dynamics of Phase Synchronization in a Family of Coupled Several One-Dimensional Chaotic Maps ................................................................. 531
Masahiro Wada (Konan University), Yoshifumi Nishio (Tokushima University)

08:50 – 09:10 Supertransients in Coupled Map Lattice with Distributed Feedback Control ............ 535
Keiji Konishi (Osaka Prefecture University)

09:10 – 09:30 Stochastic Resonance in Coupled Costas Loops with Imperfect Phase Detectors .... 539
Hisato Fujisaka (Hiroshima City University), Asuka Watanabe (Hiroshima City University), Takeshi Kamio (Hiroshima City University)

09:30 – 09:50 Asymmetrical Chaotic Coupled System Using Two Different Parameter Sets .......... 543
Yasuteru Hosokawa (Shikoku University), Yoshifumi Nishio (Tokushima University)

09:50 – 10:10 Parasitic Capacitances and Chaotic Dynamics in a Log-domain Oscillator .......... 547
Alon Ascoli (University College Dublin), Orla Feely (University College Dublin), Paul Curran (University College Dublin)

SPECIAL SESSION

SS22 — Complex Networks Applications .......................................................... 551

Date: Sep, 13th  Time: 10:30 – 12:10   Room: Imperiale A
Organized by: Francis C.M. Lau (Hong Kong Polytechnic University), Chi K. Tse (Hong Kong Polytechnic University)
Chaired by: Francis Lau (Hong Kong Polytechnic University), Chi K. Tse (Hong Kong Polytechnic University)

10:30 – 10:50 Packet Traffic Analysis in a Complex Network with Multiple Power-Law Properties ...... 551
Yongxiang Xia (Hong Kong Polytechnic University), Chi K. Tse (Hong Kong Polytechnic University), Francis C.M. Lau (Hong Kong Polytechnic University)

10:50 – 11:10 State-feedback Control of Complex Dynamical Networks ........................................ 555
Zhengping Fan (City University of Hong Kong), Francesco Sorrentino (University of Naples Federico II), Guanrong Chen (City University of Hong Kong)

11:10 – 11:30 Growing Self-Organizing Maps for Network Construction .................................... 559
Toshimichi Saito (Hosei University), Takeshi Ehara (Hosei University), Satoshi Suzuki (Hosei University), Hiroyuki Torikai (Hosei University)

11:30 – 11:50 Study of LDPC Codes Built on Scale-Free Networks ............................................ 563
Xia Zheng (Hong Kong Polytechnic University), Francis C.M. Lau (Hong Kong Polytechnic University), Chi K. Tse (Hong Kong Polytechnic University)

11:50 – 12:10 Graph-based Criteria for Synchronization of Diffusively Coupled Oscillators .......... 567
Igor Belykh (Georgia State University), Martin Hasler (EPFL), Vladimir Belykh (Volga State Academy)
SPECIAL SESSION
SS24 — Multidimensional Mobile Information Networks – II .......................... 571

Date:  Sep, 13th  Time:  10:30 – 12:10  Room:  Imperiale B
Organized by:  Shoji Shinoda (Chuo University)
Chaired by:  Keisuke Nakano (Niigata University)

10:30 – 10:50  Tracking Cost of Mobile Agents in Ad Hoc Networks .......................... 571
Ryo Nakanishi (Gunma University), Shota Kawakami (Gunma University), Yoshikuni Onozato (Gunma University)

10:50 – 11:10  Location-aided Energy-aware Routing Method for UWB-based Sensor Networks .......................... 575
Kyungsup Kwak (Inha University), Xizhi An (Inha University)

11:10 – 11:30  Context Acquisition via Multifarious Sensors in a Real World Home Ubiquitous Environment .......................... 579
Tatsuya Yamazaki (NICT)

11:30 – 11:50  Chaos-Based Generation of Optimal Spreading Sequences for DS-UWB Sensor Networks .......................... 583
Giampaolo Cimatti (University of Bologna), Riccardo Rovatti (University of Bologna), Gianluca Setti (University of Ferrara)

11:50 – 12:10  Formulation of an Assignment Problem in Wireless LANs .......................... 587
Hiroshi Tamura (Niigata Institute of Technology), Masakazu Sengoku (Niigata University), Shoji Shinoda (Chuo University)

SPECIAL SESSION
SS16 — Neurodynamics and its Hardware Implementation – II .......................... 591

Date:  Sep, 13th  Time:  10:30 – 12:10  Room:  Saturno A
Organized by:  Koji Nakajima (Tohoku University)
Chaired by:  Hatsuo Hayashi (Kyushu Institute of Technology), Yoshihiro Hayakawa (Tohoku University)

10:30 – 10:50  Slow Spiking and Variability of Inter-Spike Intervals near the Hopf Bifurcation in Simple Single Neuron Models .......................... 591
Junko Inoue (Kyoto Koka Women's University), Shinji Doi (Osaka University), Sadatoshi Kumagai (Osaka University)

10:50 – 11:10  Theta Phase Coding in Entorhinal-Hippocampal Loop Circuits Organized by Afferent Signals .......................... 595
Jun Igarashi (Kyushu Institute of Technology), Hatsuo Hayashi (Kyushu Institute of Technology), Katsumi Tateno (Kyushu Institute of Technology)

11:10 – 11:30  A Hierarchical Brain Inspired Computing System .......................... 599
Christopher Johansson (Royal Institute of Technology), Anders Lansner (Royal Institute of Technology)

11:30 – 11:50  Optimization at Work in the Cortex: Limit Cycle Computation, Minimization of the Total Wiring Length, and Architectural Computational Classes .......................... 603
Ruedi Stoop (Institute of Neuroinformatics ETHZ), Stefan Martignoli (Institute of Neuroinformatics ETHZ), Clemens Wagner (Institute of Neuroinformatics ETHZ)

11:50 – 12:10  The Nontrivial Grammar of Drosophila Courtship Language Distinguishes among Different Genetic Profiles .......................... 607
Ruedi Stoop (Institute of Neuroinformatics ETHZ), Christian Heid (Institute of Neuroinformatics ETHZ), Benjamin I Arthur (Institute of Neuroinformatics ETHZ)
Neural Networks – III

Date: Sep, 13th  Time: 10:30 – 12:10  Room: Giove A
Chaired by: Kenya Jin’no (Kanto Gakuin University), Petr Lansky (Academy of Sciences of the Czech Republic)

10:30 – 10:50 Characteristic of Hysteresis Oscillatory Associative Memory

Kenya Jin’no (Kanto Gakuin University)

10:50 – 11:10 Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition

Yuko Osana (Tokyo University of Technology)

11:10 – 11:30 The Parameters of the Ornstein-Uhlenbeck Neuronal Model

Petr Lansky (Academy of Sciences of the Czech Republic), Pavel Sanda (Academy of Sciences of the Czech Republic), Jufang He (The Hong Kong Polytechnic University)

11:30 – 11:50 Oscillation Mechanism in Cyclic Coupled Neurons

Hiroyuki Kitajima (Kagawa University), Takashi Ishii (Kagawa University), Tetsuo Hattori (Kagawa University)

11:50 – 12:10 An Autoassociation Model based on Entropy Functionals

Masahiro Nakagawa (Nagaoka University of Technology)

Analysis of Oscillators

Date: Sep, 13th  Time: 10:30 – 12:10  Room: Giove B
Chaired by: Peter Kennedy (University College Cork), Yoshihiko Susuki (Kyoto University)

10:30 – 10:50 Analysis of Phase Noise and the Effect of Large Signal Nonlinearity in Dual Cross-Coupled LC Oscillators

Sohrab Samadian (UCLA)

10:50 – 11:10 3 GHz Spread Spectrum Clock Generator for Serial ATA-II using Random Frequency Modulation

Luca A. De Michele (University of Bologna), Fabio Pareschi (University of Bologna / University of Ferrara), Riccardo Rovatti (University of Bologna), Gianluca Setti (University of Ferrara / University of Bologna)

11:10 – 11:30 Mathematical Analysis of Injection-Locked Frequency Dividers

Tao Xu (University College Cork), Zhipeng Ye (University College Cork), Michael P. Kennedy (University College Cork)

11:30 – 11:50 Analysis and Evaluation of Harmonic Distortion in the Hartley Oscillator

Melita Pennisi (University of Catania), Gaetano Palumbo (University of Catania), Salvatore Pennisi (University of Catania)

11:50 – 12:10 Amplitude Response Curves of Frequency-locked Rotations

Yoshihiko Susuki (Kyoto University), Yoshihiko Ueda (Future University-Hakodate)
SPECIAL SESSION

SS25 — Statistical Physics of Complex Networks ............................... 651

Date:  Sep, 13th  Time:  14:00 – 16:00  Room:  Imperiale A
Organized by:  Ljupco Kocarev (University of California, San Diego), Jose Amigo (Universidad Miguel Hernandez)
Chaired by:  Ljupco Kocarev (University of California, San Diego), Jose Amigo (Universidad Miguel Hernandez)

14:00 – 14:20  Statistical Physics in Information Theory: An Overview ............................... 651
Ljupco Kocarev (UCSD), José M. Amigó (Universidad Miguel Hernandez)

14:20 – 14:40  Statistical Physics of Information Flow in Wireless Networks ............................... 655
Massimo Franceschetti (UC San Diego)

14:40 – 15:00  Network Structural Properties, Communication Models and Traffic Dynamics ............................... 659
Mario Di Bernardo (University of Bristol / University of Naples Federico II), Francesco Sorrentino (University of Naples Federico II)

15:00 – 15:20  Queuing Networks with Periodic Fluid Model ............................... 663
Tommaso Addabbo (University of Siena), Ljupco Kocarev (University of California, San Diego)

15:20 – 15:40  Adaptive Traffic in Complex Networks ............................... 667
Attila Fekete (Collegium Budapest), Gabor Vattay (Collegium Budapest), Ljupco Kocarev (University of California, San Diego)

15:40 – 16:00  Information Processing in Networks of Coupled Hindmarsh-Rose Neurons ............................... 671
Paolo Checco (Polytechnic of Torino), Marco Rignero (Polytechnic of Torino), Mario Biey (Polytechnic of Torino), Ljupco Kocarev (University of California, San Diego)

Discrete Time Chaos ............................... 675

Date:  Sep, 13th  Time:  14:00 – 16:00  Room:  Imperiale B
Chaired by:  David Arrowsmith (University of London), Antony J. Lawrance (University of Warwick)

14:00 – 14:20  Queues, Long-range Dependent Traffic and Non-linear Maps ............................... 675
Raul J. Mondragon (University of London), David K. Arrowsmith (University of London)

14:20 – 14:40  An autonomous Piecewise-Constant System with Multirate Sample-Hold Controllers ............................... 679
Tadashi Tsubone (Nagaoka University of Technology), Yasuhiro Wada (Nagaoka University of Technology)

14:40 – 15:00  Fractal Structures of Three-Segment Piecewise-Linear Map in Reverse Time ............................... 683
Alexey I. Tomashevskiy (Moscow Power Engineering Institute (Technical University)

15:00 – 15:20  Applications of High-Dimensional Cat Map and Its Mixing Nature ............................... 687
H. S. Kwok (City University of Hong Kong), Wallace Tang (City University of Hong Kong)

15:20 – 15:40  On Some Properties of Chaotic Maps Implemented in Finite-State Machines ............................... 691
Mieczysław Jessa (Poznan University of Technology)

15:40 – 16:00  Higher Order Dependency of Chaotic Maps ............................... 695
Antony J. Lawrance (University of Warwick), Theodore Papamarkou (University of Warwick)
**SPECIAL SESSION**

**SS21 — Nonlinear Speech and Image Processing** .................................................. 699

**Date:** Sep, 13th **Time:** 14:00 – 16:00 **Room:** Saturno A

**Organized by:** Makoto Nakashizuka (Osaka University)

**Chaired by:** Makoto Nakashizuka (Osaka University), Akira Asano (Hiroshima University)

**14:00 – 14:20** A Short Duration Noise Suppression Method for Speech Signals by Using a Sparse Signal Representation ................................................................. 699
Makoto Nakashizuka (Osaka University)

**14:20 – 14:40** A Noise Variance Estimator for Speech Enhancement based on MAP Spectral Amplitude Estimation ................................................................. 703
Arata Kawamura (Osaka University), Youji Iiguni (Osaka University)

**14:40 – 15:00** Soliton Equations of Connection Energy for Speech Spectral Interpolations ............... 707
Yoshinao Shiraki (Shonan Institute of Technology)

**15:00 – 15:20** Unsupervised Structuring Element Optimization of Morphological Opening for Texture Images ............................................................... 711
Atsusi Fujiki (Kansai University), Akira Asano (Hiroshima University), Mitsuji Muneyasu (Kansai University)

**15:20 – 15:40** A Fast Binary Morphological Operator for Real Time Sequence Analysis ............... 715
Alessandro Bevilacqua (University of Bologna), Alessandro Gherardi (University of Bologna)

**15:40 – 16:00** Combining Correlation Kernels and Kansei Information in Kernel CCA in Texture Classification ............................................................... 719
Yo Horikawa (Kagawa University), Yujiro Ohnishi (Kagawa University)

**Cellular Neural Networks** ................................................................. 723

**Date:** Sep, 13th **Time:** 14:00 – 16:00 **Room:** Giove A

**Chaired by:** Norikazu Takahashi (Kyushu University), Marco Gilli (Polytechnic of Torino)

**14:00 – 14:20** Comparing Two-Layer CNN with van der Pol Oscillators Coupled by Inductors ........... 723
Koji Urabe (Niihama National College of Technology), Yoshifumi Nishio (Tokushima University)

**14:20 – 14:40** Study on the Number of Equilibrium Points for a 1-D DBCNN with No Input ............... 727
Hajime Hara (Hiroshima Institute of Technology), Tetsuo Nishi (Waseda University), Norikazu Takahashi (Kyushu University)

**14:40 – 15:00** Prediction of Chaos Attractor by using DT-CNN ........................................... 731
Masatomo Inoue (Sophia University), Yuta Watanabe (Fuji Xerox Co., Ltd.), Tatsuro Shimada (Toyota Motor Corporation), Masayuki Yamauchi (Hiroshima Institute of Technology), Mamoru Tanaka (Sophia University)

**15:00 – 15:20** Further Analysis on Condition for 1-D CNNs to Perform Connected Component Detection ............................................................... 735
Norikazu Takahashi (Kyushu University), Tetsuo Nishi (Waseda University)

**15:20 – 15:40** Pattern Formation in Locally Coupled 1D Arrays of Square-wave Generator Cells ........... 739
Mustak E Yalcin (Istanbul Technical University)

**15:40 – 16:00** Global Optimization with Coupled Local Minimizers Excited by Gaussian White Noise .... 743
Special Session

SS19 — Rhythmic and Synchronous Activities in Biological Systems .............. 747

Date: Sep, 13th  Time: 14:00 – 16:00  Room: Giove B

Organized by: Kunichika Tsumoto (ERATO JST/The University of Tokyo), Shigeki Tsuji (ERATO JST/The University of Tokyo), Kazuyuki Aihara (ERATO JST/The University of Tokyo)

Chaired by: Kunichika Tsumoto (The University of Tokyo/ERATO JST), Shigeki Tsuji (The University of Tokyo/ERATO JST)

14:00 – 14:20  Dependence of Synchronous Activities in Coupled Inhibitory Neurons upon Excitability Classes ......................................................... 747

Shigeki Tsuji (The University of Tokyo/ERATO JST), Tetsushi Ueta (Tokushima University), Hiroshi Kawakami (Tokushima University), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

14:20 – 14:40  Synchronous Activities in Neuronal Networks Composed of Homogeneous and Heterogeneous Neurons .......................................................... 751

Kunichika Tsumoto (The University of Tokyo/ERATO JST), Hiroyuki Kitajima (Kagawa University), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

14:40 – 15:00  A Study on a Neural Network Model of Theta Rhythm and Sequential Memory .............. 755

Youhei Nakamura (The University of Tokyo), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

15:00 – 15:20  Entrainment of Circadian Oscillations under 24h Light-Dark Cycles .......................... 759

Gen Kurosawa (The University of Tokyo/ERATO JST), Kunichika Tsumoto (The University of Tokyo/ERATO JST), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

15:20 – 15:40  Dynamic Transition among Memories on Neurocomputer Composed of Amoeboid Cell with Optical Feedback ................................................. 763

Masashi Aono (RIKEN (The Institute of Physical and Chemical Research), Masahiko Hara (RIKEN (The Institute of Physical and Chemical Research)

15:40 – 16:00  Synchronization Experimentally Observed in Calling Behaviors of Japanese Rain Frogs (Hyla-japonica) ......................................................... 767

Ikkyu Aihara (Kyoto University), Hiroyuki Kitahata (Kyoto University), Shunsuke Horai (The University of Tokyo/ERATO JST), Kazuyuki Aihara (The University of Tokyo/ERATO JST), Kenichi Yoshikawa (Kyoto University)

Analysis of Dynamical and Uncertain Systems .............................................. 771

Date: Sep, 13th  Time: 16:20 – 18:20  Room: Imperiale A

Chaired by:  Yasunori Endo (University of Tsukuba), Yuchi Kanzawa (Shibaura Institute of Technology)

16:20 – 16:40  Measurement Noise Reduction by Local Projection Methods ................................. 771

Maria E. Mera (Universidad Complutense de Madrid), Manuel Morán (Universidad Complutense de Madrid)

16:40 – 17:00  A Coalgebraic Approach to Behavioral Inheritance in Dynamical Systems ............. 775

Cristian Secchi (University of Modena and Reggio Emilia), Marcello Bonfè (University of Ferrara), Cesare Fantuzzi (University of Modena and Reggio Emilia)

17:00 – 17:20  A General SPICE Model of the Stochastic Resonance Process ............................... 779

Gabriel Vasilescu (LISIF, University P. & M. Curie), Luonan Chen (Osaka Sangyo University)

17:20 – 17:40  A New Approach to Predict the Possible Bit Streams of a Sigma Delta Modulator ...... 783

Emer M. Condon (University College Dublin), Paul Curran (University College Dublin), Orla Feely (University College Dublin)
17:40 – 18:00 Two Types of Fuzzy c-Means for Data with Tolerance
Yuchi Kanzawa (Shibaura Institute of Technology), Yasunori Endo (University of Tsukuba), Sadaaki Miyamoto (University of Tsukuba)

18:00 – 18:20 Improvement of Fuzzy ARTMAP by Controlling Match Tracking
Takeshi Kamio (Hiroshima City University), Kentaro Nomura (Hiroshima City University), Kenji Mori (Hiroshima City University), Hisato Fujisaka (Hiroshima City University), Kazuhisa Haeiwa (Hiroshima City University)

Chaos Communication
Date: Sep, 13th Time: 16:20 – 18:20 Room: Imperiale B
Chaired by: Yutaka Jitsumatsu (Kyushu University), Geza Kolumban (Budapest University of Technology and Economics)

16:20 – 16:40 Non-autonomous Chaotic Radio Pulse Generator
Nikolay Atanov (Russian Academy of Sciences), Alexander Dmitriev (Russian Academy of Sciences), Elena Efremova (Russian Academy of Sciences), Lev Kuzmin (Russian Academy of Sciences)

16:40 – 17:00 Detection of Information Symbols and Sequence Lengths Using Suboptimal Receiver for Chaos Shift Keying
Shintaro Arai (Tokushima University), Yosifumi Nishio (Tokushima University)

17:00 – 17:20 FIR Pulse Shaping Filter Makes Markov Codes Better than I.I.D. in Chip-Asynchronous DS/CDMA Systems
Yutaka Jitsumatsu (Kyushu University), Masato Ogata (Kyushu Sangyo University), Tohru Kohda (Kyushu University)

Stefano Vitali (University of Bologna), Riccardo Rovatti (University of Bologna), Gianluca Setti (University of Ferrara)

17:40 – 18:00 An M-ary Orthogonal-CSK communication system
Pascal Chargé (INSA - LESIA), Danièle Fournier-Prunaret (INSA - LESIA)

18:00 – 18:20 A Connection Between Chaotic Message-embedding and Conventional Self-synchronizing Stream
Gilles Millerioux (Henri Poincare University), José M. Amigó (Universidad Miguel Hernandez), Jamal Daafouz (Institut National Polytechnique de Lorraine)

Image Processing
Date: Sep, 13th Time: 16:20 – 18:20 Room: Saturno A
Chaired by: Mitsunori Makino (Chuo University), Noritaka Shigei (Kagoshima University)

16:20 – 16:40 Lossless Parallel Image Coding using Discrete Time Cellular Neural Networks
Ryunosuke Okamura (Sophia University), Hisashi Aomori (Sophia University), Tsuyoshi Otake (Tamagawa University), Nobuaki Takahashi (IBM Japan), Mamoru Tanaka (Sophia University)

16:40 – 17:00 Image Resolution Enhancement using Discrete-Time Cellular Neural Network
Takefumi Konishi (Tamagawa University), Hisashi Aomori (Sophia University), Tsuyoshi Otake (Tamagawa University), Nobuaki Takahashi (IBM Japan), Seiya Masuda (Tamagawa University), Mamoru Tanaka (Sophia University)

17:00 – 17:20 A Novel Edge Detection Algorithm using Cellular Neural Networks
Giuseppe Grassi (University of Lecce), Pietro Vecchio (University of Lecce), Eugenio Di Sciascio (Politechnic of Bari), Luigi A. Grieco (Politechnic of Bari), Donato Cafagna (University of Lecce)
17:20 – 17:40 Measuring the Skin Surface Changes Due to Hydrating Treatments Through Capacitive Images Analysis ................................................................. 831
   Alessandro Bevilacqua (University of Bologna), Alessandro Gherardi (University of Bologna)

17:40 – 18:00 Learning Algorithm for Multiple Vector Quantization Based Image Compression ......... 835
   Noritaka Shigei (Kagoshima University), Hiromi Miyajima (Kagoshima University), Michiharu Maeda (Kurume National College of Technology)

18:00 – 18:20 An Efficiently Adaptive Ray Tracing by Automatic Differentiation ............................. 839
   Masato Misumi (Chuo University), Mitsunori Makino (Chuo University)

Theory and Applications of Signal Processing .............................................................................. 843
Date:    Sep, 13th Time:    16:20 – 18:20 Room:    Giove A
Chaired by:    Shin Mizutani (NTT Corporation), Akira Ohsumi (Kyoto Institute of Technology)

16:20 – 16:40 Noise-assisted Quantization ................................................................. 843
   Shin Mizutani (NTT Corporation), Jun Muramatsu (NTT Corporation), Kenichi Arai (NTT Corporation), Peter Davis (NTT Corporation)

16:40 – 17:00 A Study on the Method for Evaluation of Pitch Variation of Baby’s Cry ...................... 847
   Hidenori Tanaka (Muroran Institute of Technology), Tomomasa Nagashima (Muroran Institute of Technology)

17:00 – 17:20 Nonlinear Parametric Estimation for Signals in Nonstationary Random Noise via Stationarization and Wigner Distribution ......................................................... 851
   Hiroshi Ijima (Kyoto Institute of Technology), Akira Ohsumi (Kyoto Institute of Technology), Satoshi Yamaguchi (Kyoto Institute of Technology)

17:20 – 17:40 An Iterative Warping Algorithm for Arbitrary Frequency Maps ......................... 855
   Salvatore Caporale (University of Bologna), Luca De Marchi (University of Bologna), Nicolò Speciale (University of Bologna)

17:40 – 18:00 T-entropy of EEG/EOG Sensitive to Sleep State .................................................. 859
   Mark R. Titchener (University of Auckland)

18:00 – 18:20 Feature Extraction of New Imagery Tasks Using Time-Frequency and Statistical Analysis for Brain-Computer Interface ......................................................... 863
   Montri Phothisophonthai (Nagaoka University of Technology), Masahiro Nakagawa (Nagaoka University of Technology)

Optimization ............................................................................................................................ 867
Date:    Sep, 13th Time:    16:20 – 18:20 Room:    Giove B
Chaired by:    Thomas Schimming (OCTAX Microscience GmbH), Keiji Tatsumi (Osaka University)

16:20 – 16:40 The Effect of the Parameter of the Inverse Function Delayed Model on the Success Rate of the N-Queen Problem ................................................................. 867
   Akari Sato (Tohoku University), Yoshihiro Hayakawa (Tohoku University), Koji Nakajima (Tohoku University)

16:40 – 17:00 A New Three-level Tree Data Structure for Optimizing the Traveling Salesman Problem by Lin-Kernighan Heuristic .............................................................. 871
   Hung D. Nguyen (University of Miyazaki), Ikuo Yoshihara (University of Miyazaki), Kunihito Yamamori (University of Miyazaki), Moritoshi Yasunaga (University of Tsukuba)
17:00 – 17:20 A Global Minimization Method for Separable Multimodal Functions Using a One-Dimensional Global Search .......................................................... 875
Hideo Kanemitsu (Hokkaido University of Education), Tao Feng (Fuji Electric Information Service Co., Ltd.), Hideaki Konno (Hokkaido University of Education), Masaaki Miyakoshi (Hokkaido University), Masaru Shimbo (Hokkaido Information University)

17:20 – 17:40 Cooperative Behavior in Coupled Simulated Annealing Processes with Variance Control .......................................................... 879

17:40 – 18:00 Metaheuristic Method by using a Chaos Generator with Sinusoidal Perturbations for Global Optimization .......................................................... 883
Keiji Tatsumi (Osaka University), Yoshinori Obita (Osaka University), Tetsuzo Tanino (Osaka University)

18:00 – 18:20 Metaheuristic Algorithms for Container Loading Problem Using Grouping Objects ....... 887
Yasunori Endo (University of Tsukuba), Sadaaki Miyamoto (University of Tsukuba), Koki Hanzawa (University of Tsukuba), Yukihiro Hamasuna (University of Tsukuba)

Day IV, September 14th 2006

Control .......................................................... 891

Date: Sep, 14th  Time: 08:30 – 10:10  Room: Imperiale A
Chaired by: Yuzo Ohta (Kobe University), Kazuo Komatsu (Kumamoto National College of Technology)

08:30 – 08:50 On-line Reference Inputs Management for Constrained Servo Systems ............... 891
Yuzo Ohta (Kobe University), Takaaki Taguchi (Kobe University), Tetsuya Yamaguchi (Kobe University)

08:50 – 09:10 A New Absolute Stability Criterion for 2-D Single Variable Lur’e Systems ........... 895
Róisín Duignan (University College Dublin), Paul Curran (University College Dublin)

09:10 – 09:30 On the Modeling of Nonlinear Multi-time-scale Systems — A Recursive Order Reduction Procedure .......................................................... 899
Hajer Bouzaouache (LECAP/EPT), Nizar Rezgui (LECAP/EPT)

09:30 – 09:50 Multi-objective Design Scheme for Robust Pole Assignment .......................... 903
Takanori Tagami (Anan National College of Technology), Kenji Ikeda (The University of Tokushima)

09:50 – 10:10 Design of Formal Linearization and Observer for Time-Delay Nonlinear Systems .... 907
Kazuo Komatsu (Kumamoto National College of Technology), Hitoshi Takata (Kagoshima University)
**SPECIAL SESSION**

**SS04 — Nonlinear Dynamics in Nanosystems** .......................................................... 911

**Date:** Sep, 14th  **Time:** 08:30 – 10:10  **Room:** Imperiale B

**Organized by:** Arvind Raman (*Purdue University*), Takashi Hikihara (*Kyoto University*), Igor Mezic (*University of California, Santa Barbara*)

**Chaired by:** Takashi Hikihara (*Kyoto University*)

08:30 – 08:50  **Chaos in Dynamic Atomic Force Microscopy** .......................................................... 911

A. Raman (*Purdue University*), SQ. Hu (*Purdue University*)

08:50 – 09:10  **Intermittency in Dynamic Atomic Force Microscopy** .......................................................... 915

Ferdinand Jamitzky (*Max-Planck-Institut f. Extraterrestrische Physik / Ludwig-Maximilians-Universität*),
Robert Stark (*Ludwig-Maximilians-Universität*)

09:10 – 09:30  **Energy of Floating Particles Aggregates** .......................................................... 919

Vakhtang Putkaradze (*Colorado State University*)

09:30 – 09:50  **Traveling Localized Excitations in Micro-cantilever Arrays** .......................................................... 923

Masayuki Sato (*Kanazawa University*), Albert J. Sievers (*Cornell University*)

09:50 – 10:10  **Suppression of Subharmonics and Chaos in Tapping Mode Atomic Force Microscopy Using Time Delayed Feedback Control** .......................................................... 927

Kohei Yamasue (*Kyoto University*), Takashi Hikihara (*Kyoto University*)

**Mathematical Physics** .......................................................... 931

**Date:** Sep, 14th  **Time:** 08:30 – 10:10  **Room:** Saturno A

**Chaired by:** Maciej Ogorzalek (*Jagiellonian University*), Hideaki Matsueda (*Kochi University*)

08:30 – 08:50  **Lagrangian Stability of Viscous Flows** .......................................................... 931

Yodai Watanabe (*National Institute of Informatics*)

08:50 – 09:10  **Maximum Amplitude and Instability of a Shallow Water Solitary Wave** .......................................................... 935

Shinsuke Watanabe (*Yokohama National University*), M. Oikawa (*Kyushu University*), H. Tsuji (*Kyushu University*), Y. Funakubo (*Yokohama National University*), K. Ohtsuka (*Yokohama National University*), T. Minagawa (*Yokohama National University*)

09:10 – 09:30  **Nonlinear Photonic Interactions among Electric Multipoles in Quantum Dots** .......................................................... 939

Hideaki Matsueda (*Kochi University*)

09:30 – 09:50  **Fast High-Dimensional Boxcounting with Ternary Search Trees** .......................................................... 943

Christian Merkwirth (*Jagiellonian University*), Maciej Ogorzalek (*Jagiellonian University*), Ulrich Parlitz (*University of Goettingen*)

09:50 – 10:10  **A Method of Proving Existence of Solution Curve for Nonlinear Equation using Affine Arithmetic** .......................................................... 947

Yuchi Kanzawa (*Shibaura Institute of Technology*), Shin’ichi Oishi (*Waseda University*)
Communications ................................................................. 951

Date:  Sep, 14th  Time:   08:30 – 10:10  Room:  Giove A
Chaired by:  Hisa-Aki Tanaka (The University of Electro-Communications), Gianluca Mazzini (University of Ferrara)

08:30 – 08:50  Evaluation of Route Expected Transmission Count of Routing Strategies in Multi-hop Wireless Networks ................................................................. 951
Kazuyuki Miyakita (Niigata University), Keisuke Nakano (Niigata University), Masakazu Sengoku (Niigata University), Shoji Shinoda (Chuo University)

08:50 – 09:10  Channel Features Observed to Model the Long Term Behavior of Wireless Links ......... 955
Eugenio Costamagna (University of Pavia), Lorenzo Favalli (University of Pavia), Pietro Savazzi (University of Pavia)

09:10 – 09:30  Self-adaptive Timing-synchronization Ability in Ad Hoc Networks based on IEEE 802.11 MAC Protocol ................................................................. 959
Hisa-Aki Tanaka (The University of Electro-Communications), Osamu Masugata (The University of Electro-Communications), Daisuke Ohta (The University of Electro-Communications), Akio Hasegawa (ATR Adaptive Communications Research Laboratories), Peter Davis (ATR Adaptive Communications Research Laboratories)

09:30 – 09:50  P3P : a P2P Overlay Multicast Network for Multimedia Streaming ......................... 963
Andrea Odorizzi (University of Ferrara), Gianluca Mazzini (University of Ferrara)

Raffaele Rugin (University of Ferrara), Gianluca Mazzini (University of Ferrara)

Work in Progress – I ............................................................. 971

Date:  Sep, 14th  Time:   08:30 – 10:10  Room:  Giove B
Chaired by:  Toshimitsu Ushio (Osaka University)

08:30 – 08:50  Forced Synchronization of Class I and Class II Neurons with both electrical and Chemical Synapses ................................................................. 971
Eri Ioka (Kagawa University), Hiroyuki Kitajima (Kagawa University)

08:50 – 09:10  Critical Temperature Sensor Based on Spiking Neuron Models ......................... 975
Gessyca M. Tovar (Hokkaido University), Tetsuya Hirose (Hokkaido University), Tetsuya Asai (Hokkaido University), Yoshihito Amemiya (Hokkaido University)

09:10 – 09:30  Consensus Problem of Multi-agent Systems with Non-linear Performance Functions ... 979
Naoki Hayashi (Osaka University), Toshimitsu Ushio (Osaka University), Fumiko Harada (Osaka University), Atsuko Ohno (Osaka University)

09:30 – 09:50  Learning Partial Synchronization Regimes with Imposed Qualitative Behavior on an Array of Chua’s Oscillators ................................................................. 983

09:50 – 10:10  Global Noise Estimation Based on Tensor Product Expansion with Absolute Error .... 987
Akitoshi Itai (Aichi Prefectural University), Hiroshi Yasukawa (Aichi Prefectural University), Ichik Takumi (Nagoya Institute of Technology), Masayasu Hata (Chubu University)
**SPECIAL SESSION**

**SS20 — Advanced Modeling and Simulations Techniques**

- **Date:** Sep, 14th  
- **Time:** 10:30 – 12:10  
- **Room:** Imperiale A  
- **Organized by:** Hideki Asai *(Shizuoka University)*  
- **Chaired by:** Yuichi Tanji *(Kagawa University)*  

10:30 – 10:50 Comparison between PSpice Simulation, PSIM Simulation and their Co-Simulation with Simulink for Transient Analysis of Three-Phase Nonlinear Circuits

Giuseppe Acciani *(Polytechnic of Bari)*, Francesco Vacca *(Polytechnic of Bari)*, Silvano Vergura *(Polytechnic of Bari)*

10:50 – 11:10 Reduced Bifurcation Diagram for Harmonic Balance Method using Invariant

Masakazu Yagi *(Kyoto University)*, Takashi Hisakado *(Kyoto University)*

11:10 – 11:30 On Optimization Algorithm for Attaining the Maximum DC Gain of CMOS Amplifiers

Masayoshi Oda *(Tokushima University)*, Yoshihiro Yamagami *(Tokushima University)*, Yoshifumi Nishio *(Tokushima University)*, Junji Kawata *(Tokushima Bunri University)*, Akio Ushida *(Tokushima Bunri University)*

11:30 – 11:50 Modeling and Simulation of Delta-Sigma Fractional-N PLL Frequency Synthesizer in Verilog-AMS

Zhipeng Ye *(University College Cork)*, Michael P. Kennedy *(University College Cork)*

11:50 – 12:10 RLCG-MNA Formulation for Fast Simulation and MOR of RLC Networks

Yuichi Tanji *(Kagawa University)*, Takayuki Watanabe *(University of Shizuoka)*, Hideki Asai *(Shizuoka University)*

---

**SPECIAL SESSION**

**SS18 — Nonlinear Information Processing in the Auditory Feedback Loop**

- **Date:** Sep, 14th  
- **Time:** 10:30 – 12:10  
- **Room:** Imperiale B  
- **Organized by:** Ruedi Stoop *(Swiss Federal Institute of Technology Zurich)*  
- **Chaired by:** Ruedi Stoop *(Swiss Federal Institute of Technology Zurich)*  

10:30 – 10:50 Towards a Quantitative Theory of Biocomputation

Ruedi Stoop *(Institute of Neuroinformatics ETHZ)*, Thomas Ott *(Institute of Neuroinformatics ETHZ)*

10:50 – 11:10 Response of the Hopf Cochlea to Complex Input Sounds

Stefan Martignoli *(Institute of Neuroinformatics ETHZ)*, Albert Kern *(Institute of Neuroinformatics ETHZ)*, Ruedi Stoop *(Institute of Neuroinformatics ETHZ)*

11:10 – 11:30 Acoustic Source Separation by Atomic Signal Decomposition

Albert Kern *(Institute of Neuroinformatics ETHZ)*, Thomas Ott *(Institute of Neuroinformatics ETHZ)*, Ruedi Stoop *(Institute of Neuroinformatics ETHZ)*


Norbert Stoop *(Institute of Neuroinformatics ETHZ)*, Thomas Ott *(Institute of Neuroinformatics ETHZ)*, Ruedi Stoop *(Institute of Neuroinformatics ETHZ)*

11:50 – 12:10 Faster Spike Sorting with Belief Propagation

Thomas Ott *(Institute of Neuroinformatics ETHZ)*, Albert Kern *(Institute of Neuroinformatics ETHZ)*, Ruedi Stoop *(Institute of Neuroinformatics ETHZ)*
Chaos and Bifurcation

Date: Sep, 14th  Time: 10:30 – 12:10  Room: Saturno A
Chaired by: Federico Bizzarri (University of Genova), Hiroyuki Kitajima (Kagawa University)

10:30 – 10:50 Bifurcation Structure of Coexisting Two Duck Solutions and their Breakdown into Chaos
Munehisa Sekikawa (ERATO JST), Naohiko Inaba (Utsunomiya University), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

10:50 – 11:10 Analysis of Co-existence Phenomena of Superstable Periodic Orbit and Chaos in a Non-autonomous Piecewise Constant Circuit
Yusuke Matsuoka (Hosei University), Toshimichi Saito (Hosei University)

11:10 – 11:30 Codimension-2 Bifurcation Points Organizing the Bifurcation Scenario of a Hysteresis Oscillator
Federico Bizzarri (University of Genova), Oscar De Feo (University College Cork), Daniele Stellardo (University of Genova), Marco Storace (University of Genova)

11:30 – 11:50 Study of Hyperchaotic Multi-scroll Attractors via Adomian Decomposition with Application to Synchronization and Control
Donato Cafagna (University of Lecce), Giuseppe Grassi (University of Lecce)

11:50 – 12:10 Generating Chaos with a Linear Filter
Ned J. Corron (U. S. Army RDECOM), Scott T. Hayes (U. S. Army RDECOM), Shawn D. Pethel (U. S. Army RDECOM), Jonathan N. Blakely (U. S. Army RDECOM)

Work in Progress – II

Date: Sep, 14th  Time: 10:30 – 12:10  Room: Giove A
Chaired by: Sadayuki Murashima (Kagoshima University)

10:30 – 10:50 Model-Following Control for Nonlinear Plants with Time Delay
Yoichi Ishikawa (Meiji University), Yoshihisa Ishida (Meiji University)

10:50 – 11:10 Lossless Image Reconstruction by Contractive Mapping of Neighboring Pixels
Akira Hiramatsu (Kagoshima University), Korakot Prachumrak (King Mongkut's Institute of Technology), Takayasu Fuchida (Kagoshima University), Sadayuki Murashima (Kagoshima University)

11:10 – 11:30 A Distributed Hybrid Optimization System for NDET Inverse Problems
Anton Duca (Polytechnic University Bucharest), Florian Tomescu (Polytechnic University Bucharest)

11:30 – 11:50 Sleep Mode Investigation in Ad Hoc Networks
Alberto Bisello (University of Ferrara), Alessandra Giovanardi (University of Ferrara), Gianluca Mazzini (University of Ferrara)
Work in Progress – III
Date: Sep, 14th  Time: 10:30 – 12:10  Room: Giove B
Chaired by: Ichiro Kanaya (Osaka University)

10:30 – 10:50  IC Implementations of Shinriki’s and Inaba’s Chaotic Circuits
Takuya Hamada (Tokyo Denki University), Munehisa Sekikawa (ERATO JST), Yoshikko Horio (Tokyo Denki University), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

Bacha Anis (Laboratoire d’Etude et Commande Automatique des Processus), Jerbi Houssem (Faculté des Sciences de Sfax), Benhadji B. Naceur (Laboratoire d’Etude et Commande des Processus)

11:10 – 11:30  Synthesis and Analysis of Aesthetic Fractal Patterns for Textile Design
Takeshi Sakamoto (Osaka University), Ichiro Kanaya (Osaka University), Kosuke Sato (Osaka University)

11:30 – 11:50  Improved Parallel Processing Hardware Algorithm for Large-Scale Quadratic Assignment Problems
Naoki Ogawa (Tokyo Denki University), Yuya Ohashi (Tokyo Denki University), Yoshikko Horio (Tokyo Denki University), Kazuyuki Aihara (The University of Tokyo/ERATO JST)

PLENARY SESSION – III
Date: Sep, 14th  Time: 14:00 – 15:00  Room: Imperiale A + B
Chaired by: Takashi Hisakado (Kyoto University)

14:00 – 15:00  Cycles in evolutionary game theory
Karl Sigmund (University of Vienna)

PLENARY SESSION – IV
Date: Sep, 14th  Time: 15:30 – 16:30  Room: Imperiale A + B
Chaired by: Riccardo Rovatti (University of Bologna)

15:30 – 16:30  Locally Complex and Globally Simple: the Goal of Evolution in Biological Networks
Sergio Rinaldi (Polytechnic of Milan)
Conference Papers
This page has been intentionally left blank.
Synchronization Regions and Behaviour of Nonlinear Periodic Circuits

Giancarlo Storti Gajani and Angelo Brambilla

Dipartimento di Elettronica e Informazione,
Politecnico di Milano,
Email: storti@elet.polimi.it, brambill@elet.polimi.it

Abstract—Nonlinear periodic circuits, i.e. oscillators, are used in a large number of common applications for example telecommunications. Robust simulation methods, in particular methods capable of finding steady state solutions in presence of small amplitude forcing signals are of great importance and constitute the background for the determination of phase and amplitude noise in circuit simulators. Most current approaches are based on the variational model of the unperturbed oscillator. However synchronization effects can lead to incorrect or completely wrong results.

1. Introduction

The numerical computation by means of a circuit simulator of the amplitude and phase modulation signals generated when a small periodic perturbation forces an autonomous periodic circuit is of great importance for example in circuits for telecommunications. Amplitude and phase modulation spread the bandwidth of each component of the original spectra of unperturbed oscillators and the resulting spectra are of main importance, for example during the design of radio frequency equipments, since the characteristics of the local oscillator affect the performance of the entire system.

The common approach in the simulation of the steady state behaviour of nonlinear periodic circuits is based on the variational model and assumes that the forcing signal effects are additive to the steady state solution of the unperturbed circuit. We show that this choice is reasonable if the frequency of the forcing signal is far from the working frequency of the unperturbed circuit, but can lead to inaccurate or wrong results once the forcing frequency approaches that of the free-running oscillator. In this case, synchronization effects must be accounted for, and, being these phenomena intrinsically nonlinear [1], the variational model, that assumes a fixed relative phase between perturbation and system, in general fails. Two simple examples, a Van der Pol oscillator and a “polar” oscillator, are analysed showing some peculiar results and evident synchronization with respect to the forcing signal.

2. The variational model

Variational models of autonomous systems are known since a long time [4] [3]: main concepts are here recalled.

Consider the state equation

$$\frac{dy}{dt}(t) = f(y(t), \nu(t))$$

where $y(t) : \mathbb{R} \rightarrow \mathbb{R}^L$, $f(y(t), \nu(t)) : \mathbb{R}^{L+P} \rightarrow \mathbb{R}^L$, being $L$ the state dimension of the system and $\nu(t) \in \mathbb{R}^P$ a perturbation applied to the system. Assume that the corresponding unperturbed system

$$\frac{dy}{dt}(t) = f(y(t), 0)$$

admits the asymptotically stable $T$-periodic solution $\tilde{y}(t, t_0)$, i.e. $\tilde{y}(t, t_0) = \tilde{y}(t + T, t_0) \forall t$. Different initial conditions in general yield different steady state solutions; assume that each of these solutions is a time shifted version of $\tilde{y}(t, t_0)$. The perturbed solution of the system can be expressed as $y(t) = \tilde{y}(t) + y_s(t)$ where $y_s(t) \in \mathbb{R}^L$ is the vector of state variable variations.

To describe the response of the perturbed system in a neighborhood of the nominal periodic solution $\tilde{y}(t)$, state equation (1) is linearized by adopting a Taylor expansion of $f(y(t), \nu(t))$ truncated at the first term

$$f(y(t), \nu(t)) \approx f(\tilde{y}(t), 0) +$$

$$+ \frac{\partial f}{\partial y} \bigg|_{y=\tilde{y}, \nu=0} (y(t) - \tilde{y}(t)) + \frac{\partial f}{\partial \nu} \bigg|_{y=\tilde{y}, \nu=0} \nu(t)$$

Since $\tilde{y}(t)$ is a solution of (2), we obtain the linear time varying system referred to as variational model

$$\frac{dy_s}{dt}(t) = G(t) y_s(t) + B(t) \nu(t)$$

where $G(t) \in \mathbb{R}^{L \times L}$ and $B(t) \in \mathbb{R}^{L \times P}$ are

$$G(t) = \frac{\partial f}{\partial y} \bigg|_{y=\tilde{y}(t), \nu=0} , \quad B(t) = \frac{\partial f}{\partial \nu} \bigg|_{y=\tilde{y}(t), \nu=0}$$

The nonlinearity in (1) corresponds to time dependency of matrix functions $G(t)$ and $B(t)$. The linearization was obtained by adopting a precise time reference depending on initial conditions, but, if the frequency of the forcing signal (supposed sinusoidal) is close enough to the frequency of their unperturbed steady state solution and if the relative phase between the forcing signal and the oscillators is
not properly chosen, the variational equation can have little or no correspondence to the problem being analysed. In other words, given an oscillator and a periodic perturbation whose amplitude and frequency are inside the synchronization region, different variational equations must be used for perturbations having different phase values.

3. The Van der Pol oscillator

Consider a forced Van der Pol oscillator:

\[
\frac{d^2x}{dt^2}(t) + \mu \left(x^2(t) - 1\right) \frac{dx}{dt}(t) + \omega_0^2 x(t) + K \cos(\omega_s t + \phi) = 0 \tag{5}
\]

where \(K\) is the amplitude of the forcing signal, \(\omega_s\) is its angular frequency and \(\phi\) is a generic phase; search for a reasonable approximation of its steady state solution. With the transformation \(\tau = \omega t\), where \(\omega\) is the unknown angular frequency of the oscillator, we have

\[
\omega_0^2 x(t) + \omega^2 (x^2(t) - 1) \frac{dx}{d\tau}(\tau) + \omega_0^2 \tau = 0 \tag{6}
\]

where \(\beta = \omega_s/\omega\). We assume that \(\omega\) and \(x(\tau)\) can be expanded as power series with respect to \(\mu\) that represents the amount of “nonlinearity” characterizing the oscillator since \(\omega \to \omega_0\) as \(\mu \to 0\):

\[
\omega = \omega_0 + \mu \omega_1 + \mu^2 \omega_2 + \mu^3 \omega_3 + \ldots
\]

\[
x(\tau) = x_0(\tau) + \mu x_1(\tau) + \mu^2 x_2(\tau) + \ldots
\tag{7}
\]

By substituting (7) in (6) we obtain an equation (not shown) whose possible solution can be obtained by grouping the coefficients of the powers of \(\mu\) and equating them to zero (only the first two terms are shown):

\[
\mu^0 \rightarrow \omega_0^2 \frac{d^2 x_0}{d\tau^2} + \omega_0^2 x_0 + K \cos(\beta \tau + \phi) = 0 \tag{8}
\]

\[
\mu^1 \rightarrow \omega_0^2 \frac{d^2 x_1}{d\tau^2} + 2 \omega_0 \omega_1 \frac{d^2 x_0}{d\tau^2} + \omega_0 (x_0^2 - 1) \frac{dx_0}{d\tau} + \omega_0^2 x_1 = 0 \tag{9}
\]

By substituting the particular solution we have \(\zeta \cos(\beta \tau + \phi)\) in (8):

\[
\omega_0^2 (\beta^2 - 1) \zeta = K \implies \zeta = \frac{K}{\omega_0^2 (\beta^2 - 1)}
\]

showing that \(\beta\) must be different from 1, that is, the angular frequency of the forcing signal must be different from that of the oscillator. The solution of (8) is

\[
x_0(\tau) = C_0 \cos(\tau) + D_0 \sin(\tau) + \frac{K}{\omega_0^2 (\beta^2 - 1)} \cos(\beta \tau + \phi)
\tag{10}
\]

With condition \(\frac{dx}{d\tau}(0) = 0\) we have \(D_0 = 0\). Imposing this condition is equivalent to set “a time reference”. The \(\phi\) phase of the forcing signal is thus related to a specific oscillator reference and represents a “relative phase”. By substituting (10) in (9) and solving in \(x_1(\tau)\) we have \(\omega_1 = 0, C_0 = \sqrt{4 - 2 \sigma^2}\) and \(\beta \neq 3\), (i.e. \(\omega_s\) different from the third harmonic of the oscillator).

As expected, the expressions of \(x_0(\tau)\) and of \(x_1(\tau)\) show that the spectrum of the solution is composed of the two fundamentals at frequency \(\omega\) and \(\beta \omega\) respectively and by their beats. The amplitude of the \(\beta\) fundamental and that of its beats depend on \(K\) and in particular vanish for \(K \to 0\).

In the derivation of \(x_1(\tau)\) it is found that \(C_0\) exists only if \(\zeta \leq \sqrt{2}\). This is a key aspect, meaning that, if the amplitude of the driving signal is sufficiently large, the approximate steady state solution computed above is no longer valid. In the following we show that when \(K\) is sufficiently large or when \(\omega_s \to \omega\) (i.e. \(\beta \to 1\)) the oscillator synchronizes to the forcing signal and therefore the relative phase between them must assume a specific value. This aspect can not be accounted by the variational model. Recall that \(\zeta \to \infty\) when \(\beta \to 1\) meaning that for any value of \(K\) even extremely small there is a \(\omega_s\) value close enough to \(\omega\) to have synchronization. Furthermore the synchronization process is a dynamic one with its own behavior; here we consider only its steady state solution.

Substitute once more (10) in (9), and set to zero the coefficients of the \(\cos(\tau)\) and \(\sin(\tau)\) components:

\[
\begin{align*}
8 \omega_1 C_0 - D_0^2 (\beta^2 - 2) - D_0^2 C_0^2 &= 0 \\
8 \omega_1 D_0 + 2 C_0 (\beta^2 - 2) + C_0 (C_0^2 + D_0^2) &= 0
\end{align*}
\tag{11}
\]

By recalling that we have assumed \(\sqrt{2} < \zeta < \infty\), from the first equation of (11) we see that

\[
\begin{cases}
\text{if } D_0 < 0 & \text{then } \omega_1 C_0 < 0 \quad \text{condition}\ A \\
\text{if } D_0 > 0 & \text{then } \omega_1 C_0 > 0 \quad \text{condition}\ B
\end{cases}
\]

and from the second equation of (11) we see that

\[
\begin{cases}
\text{if } D_0 < 0 & \text{then } \omega_1 C_0 > 0 \quad \text{NOT condition}\ A \\
\text{if } D_0 > 0 & \text{then } \omega_1 C_0 < 0 \quad \text{NOT condition}\ B
\end{cases}
\]

Therefore \(C_0 = D_0 = 0\) is the unique admissible solution of (11). If \(\omega_s = \omega\) equations (8) and (9) are no longer valid and must be rewritten as

\[
\mu^0 \rightarrow \frac{d^2 x_0}{d\tau^2} + x_0 = 0 \tag{12}
\]

\[
\mu^1 \rightarrow \frac{d^2 x_1}{d\tau^2} + 2 \omega_1 \frac{d^2 x_0}{d\tau^2} + \frac{1}{\omega_0} (x_0^2 - 1) \frac{dx_0}{d\tau} + \frac{K}{\omega_0^2} \cos(\tau + \phi) = 0 \tag{13}
\]
If for simplicity we assume $\phi = 0$, by solving (12), substituting it in (13) and setting to zero the coefficients of $\cos(\tau)$ and $\sin(\tau)$ we obtain $\omega_1 = 0$, $C_0 = 0$ and

$$D_0 \left(D_0^2 - 4\right) = -\frac{4K}{\omega_0^2 \mu}.$$  

(14)

In this case the oscillator is synchronous and the components of $x_0(\tau)$ are orthogonal to the forcing signal, furthermore $D_0 \rightarrow 2$ for $K \rightarrow 0$. We end this section by reporting some simulation results obtained after having implemented (5) with a circuit and solved it through a circuit simulator. In all simulations $\omega_0 = 1$ and $\mu = 10^{-4}$. Waveform A shows that $C_0$ and $D_0$ in (10) are different from 0. In fact the period of this waveform is $2\pi/\omega$ while that of the forcing signal is $\pi/\phi$. On the other hand, waveform B shows that $C_0 = D_0 = 0$ since there is evidence that the period of $x(\tau)$ is equal to that of the forcing signal.

4. The “polar” oscillator

In polar coordinates a simple oscillator can be written as

$$\begin{cases}
\frac{dp}{dt} = \lambda(1 - \rho) \\
\frac{d\theta}{dt} = 2\pi
\end{cases}$$

(15)

if $\lambda > 0$, as $t \rightarrow \infty$, trajectories converge to the limit cycle, i.e. a unit circle, and the period is $T_0 = 1$. Assume now $\theta(0) \equiv \theta_0 = 0$ and $f_s \equiv \omega_s/2\pi$. By adding a sinusoidal perturbation to one of the Cartesian state components, we have in polar coordinates

$$\begin{cases}
\frac{dp}{dt} = \lambda(1 - \rho_p) + K \cos(\theta_p) \cos(2\pi f_s t + \phi) \\
\frac{d\theta}{dt} = 2\pi + \frac{K}{\rho_p} \sin(\theta_p) \cos(2\pi f_s t + \phi)
\end{cases}$$

(16)

Having set $\theta_0 = 0$, $\phi$ is the initial relative phase of the perturbation with respect to the oscillator. Unfortunately as for the Van der Pol oscillators a closed form solution is quite difficult to find, but an approximate steady state solution for the case $f_s = 1$ can be determined and has the form

$$\begin{cases}
\rho_p(t) \approx C_0 + C_1 \sin(4\pi t + C_2 + 2\phi) \\
\theta_p(t) \approx \phi + 2\pi t + D_1 \cos(4\pi t + 2\phi)
\end{cases}$$

(17)

that, given some hypotheses on the value of coefficients $C_j$ and $D_1$ satisfies with some approximation (16). The component $\theta_p(t)$ of (17) depends on the phase $\phi$ of the perturbation and the perturbed oscillator will synchronize with the perturbation except for a small periodic variation, the sinusoidal term with amplitude $D_1$, representing phase modulation. Amplitude modulation is represented by the term having amplitude $C_1$. After some computations we find:

$$\begin{align*}
C_0 &\approx 1 + \frac{K}{2\lambda} \\
C_1 &\approx \frac{K/2}{4\pi \cos C_2 + \lambda \sin C_2} \approx \frac{K/2}{4\pi} \\
C_2 &\approx \tan^{-1} \frac{\lambda}{4\pi} \\
D_1 &\approx \frac{\lambda K}{4\pi(2\lambda + K)}
\end{align*}$$

With this “highly symmetric” oscillator $D_1 \approx C_1$ for $K \rightarrow 0$ and the effects of the perturbation distributes evenly between phase and amplitude modulation. In Fig. 2 we see the behavior of relative phase $\phi$ as $f_s \rightarrow 1$, including initial transient. When the frequency of the forcing signal is very close to that of the unperturbed oscillator there is an evident synchronization effect as predicted by the approximate steady state solutions (17).

5. The synchronization region

The two examples discussed above synchronize to a perturbation if the forcing frequency is equivalent to that of
the free running oscillator. More in general it can be shown [1] that synchronization occurs for all forcing signals that satisfy a set of conditions on amplitude and frequency. The corresponding region in the amplitude vs. frequency plane is called synchronization region.

To determine the boundaries of this region, bifurcation analysis is performed on the stroboscopic map obtained by sampling the system at the same frequency as the perturbation signal. In a general framework, consider a dynamical system

$$\dot{x}(t) = F(x(t), t, T_s, K)$$  \tag{18}$$

perturbed by a periodic signal of period $T_s$ and amplitude $K$. For $K = 0$ the system is autonomous and $T_0$-periodic. The time evolution operator $\phi_t$ for (18) is described by $x(t + t_0) = \phi_t x(t_0)$ and, numerically, is equivalent to integrating $F(x)$ from $t_0$ to $t_0 + t$. The corresponding $T_s$ stroboscopic map is then defined as

$$x_{n+1} = \Phi(x_n, T_s, K) \equiv \phi_{T_s} x(t_0 + nT_s)$$  \tag{19}$$

parametric in $T_s$ and $K$ and any stable fixed point of (19) corresponds to a periodic orbit of (18) that is phase locked, i.e. synchronized, to the perturbation signal. Unfortunately, in practical cases, it is impossible to write a closed form expression for $\Phi$, but it is easy to write a wrapper function to the original system equation $F(x)$ that includes an integration routine having the endpoint of the integration interval mapped to parameter $T_s$. This method has been applied to a normalized Van Der Pol oscillator, a wrapper function that includes a standard RADAU5 integrator has been written and the continuation methods described in [2] (included in the CONTENT dynamical system simulator) have been used to simulate the system and trace the boundaries of the synchronization regions. Results of saddle-node continuation for the Van Der Pol stroboscopic map are shown in Fig. 3. Points marked with “CP” are cusps, corresponding, as expected, to points where the two branches of the synchronization boundary meet. Note that the frequency of the normalized unperturbed Van Der Pol oscillator is approximately 151 mHz. Three synchronization regions are clearly visible and correspond, from top to bottom, to 1:3, 1:2, and 1:1 resonances. The 1:2 resonance region is very small and degenerates into a single line. Considering once more the polar oscillator it is interesting to see what happens for larger values of the perturbation signal, in this case we expect to find regions where other types of bifurcations are found. Each region in Fig. 4 corresponds to a different type of synchronization behaviour, in particular in regions “a” and “d” phase is locked with real multipliers in “a” and complex multipliers in “d” (part of the boundary of “d” was not computed due to numerical convergence problems). regions “b” and “e” correspond to parameter values where no synchronization is obtained. Multipliers are real in “b” and complex in “e”.

References

Synchronization of Chaotic Circuits Linked by Cross Talk

Yuki Nakaaji and Yoshifumi Nishio

Department of Electrical and Electronic Engineering, Tokushima University
Email: \{nakaaji, nishio\}@ee.tokushima-u.ac.jp

Abstract—In this study, we investigate synchronization of two Chua’s circuits with a lossless transmission line. Two circuits with transmission lines placed in parallel cause the cross talk phenomena. The effect of the cross talk is modeled by the connections via coupling capacitors or mutual inductors.

By computer simulations, we investigate the synchronization of the two chaotic circuits linked by the cross talk. Especially, in this study, we pay our attentions to the breakdown of chaos synchronization.

1. Introduction

Recently, several researchers have investigated synchronization of chaotic circuits. Chua’s circuit is one of the simplest autonomous chaotic circuits. Chua’s circuit consists of a resistor, a capacitor, a nonlinear resistor and an LC resonator. The nonlinear resistor in Chua’s circuit has a piecewise linear \( v - i \) characteristics as shown in Fig. 1 and it is described by the following equation:

\[
i_R = m_2 v_1 + 0.5(m_0 - m_1)[|v_1 + B_{p1}| - |v_1 - B_{p1}|]
+ 0.5(m_1 - m_2)[|v_1 + B_{p2}| - |v_1 - B_{p2}|],
\]

where \( m_0, m_1 \) and \( m_2 \) are the slopes in the segments of the piecewise linear function, and \( B_{p1} \) and \( B_{p2} \) denote the breakpoints. There are many studies about coupled Chua’s circuits, for example, master-slave coupling of Chua’s circuits [1], mutual coupling of Chua’s circuits [2], a ladder of Chua’s circuits, a ring of Chua’s circuits [3], two-dimensional array of Chua’s circuits [4], etc. However, almost researchers consider the couplings by lumped elements. We have investigated chaotic phenomena observed from Chua’s circuit when the LC resonator is replaced by a transmission line [5]. Further, we have also investigated synchronization phenomena of two Chua’s circuits coupled by a transmission line [6].

We consider two Chua’s circuits with lossless transmission lines placed in parallel. Transmission lines placed in parallel cause the cross talk phenomena. Cross talk phenomena appear in long powerlines or very high speed VLSI and its effect is normally not preferable. We have already confirmed that the two Chua’s circuit with transmission lines could be synchronized by the effect of the cross talk [7]. In this study, we investigate the synchronization of the two Chua’s circuits linked by the cross talk in detail, especially, paying our attentions to the breakdown of chaos synchronization.

2. Basic Circuit Model [7]

In our previous research, we reported that two Chua’s circuits with lossless transmission lines could be synchronized. The circuit model is shown in Fig. 2.

We modeled the transmission lines by LC ladder circuits with finite numbers of lumped elements. Further, we modeled the cross talk effect by coupling capacitors or mutual inductors. The circuit equations could be derived as follows:

\[
C_i \frac{dv_j}{dt} = \frac{v_j - v_j(l-1) + C_i \frac{dv_j}{dt}}{R_{ij} + L_{ij}} + C_k \frac{dv_j(k-1) - v_j(k)}{R_{jk} + L_{jk}} \quad (j=1, 2, \ldots, n), \quad (k=1, 2, \ldots, n), \quad (l=1, 2, \ldots, n) \quad (j=1, 2). \quad \text{From Eqs. (1) and (2), we could obtain the normalized circuit equations as follows;}
\]

\[
x_{i0} &= x_{i1} - x_{i0} - f(x_{i0})
\]

\[
x_{i1} &= \alpha_1(y_{i1} - x_{i1} + x_{i10}) - \beta_{i1}(y_{i2} - x_{i2} + x_{i20})
\]

\[
x_{iK} &= \alpha_k(y_{i(k-1)} - x_{i(k-1)}) - \beta_{iK}(y_{i(k-2)} - y_{i(k-1)})
\]

\[
x_{i20} &= \zeta(x_{i21} - x_{i20} - f(x_{i20}))
\]

\[
x_{i21} &= \alpha_2(y_{i21} - \zeta(x_{i21} - x_{i21})) - \beta_{i21}(y_{i1} - x_{i1} + x_{i10})
\]

\[
x_{i2K} &= \alpha_2(y_{i(k-2)} - y_{i(k-1)}) - \beta_{i2K}(y_{i(k-1)} - y_{i(k-1)})
\]

\[
y_{ij} &= \gamma_j(x_{j(i+1)} - x_{ij}) + \delta_j(x_{j(i+1)} - x_{ij}) \quad (j=1, 2).
\]

\[
f(x_{j0}) = c_j x_{j0} + 0.5(a_j - b_j)[|x_{j0} + 1| - |x_{j0} - 1|]
+ 0.5(b_j - c_j)[|x_{j0} + d_j| - |x_{j0} - d_j|],
\]

\[
R_{ij} = R_{ij} = \frac{2}{\pi \sqrt{L_{ij} C_{ij}}}
\]

Figure 1: (a) Chua’s circuit. (b) \( v - i \) characteristics of nonlinear resistor.

Figure 2: Two Chua’s circuits with transmission lines.
\[
\begin{align*}
\tau &= R_1 \tau, \quad \alpha_{jk} = \frac{C_{jk} C_{10} (C_{j+1}k - C_k)}{-C_k - (C_{jk}^2 + 1) C_k + C_{jk}^2 C_{j+1}k}, \\
\beta_{jk} &= \frac{C_{jk} C_{10} C_k}{-C_k^2 - (C_{jk}^2 + 1) C_k + C_{jk}^2 C_{j+1}k}, \\
\beta_l &= \frac{R_{10}^2 C_{10} M_l}{(L_{2l} - M_l)(L_{1l} - M_l) - M_l^2}, \\
\gamma_{ij} &= \frac{R_{10}^2 C_{10} L_{ij}}{\zeta = C_{10}/C_{20}}, \quad \alpha_j = R_j m_0, \\
b_j &= R_j m_1, \quad c_j = R_j m_2, \quad d = \frac{B_{p2}}{B_{p1}}, \\
C_{jk} &= C_{1k}, \quad (k=2, 3, \ldots, n), \quad (l=1, 2, \ldots, n) \text{ and } (j=1, 2). 
\end{align*}
\]

Some examples of chaos synchronization obtained by computer simulations are shown in Figs. 3 and 4. We confirmed that the circuits were synchronized in anti-phase in the case of mutual inductors and in-in-phase in the case of coupling capacitors.

3. Breakdown of Synchronization

In order to understand the synchronization phenomena via cross talk effects of the transmission lines, we investigate the breakdown of the synchronization when the circuit parameters are gradually changed.

We fix the coupling parameter \( \beta \) and increase the error between the two Chua’s circuits. Figure 5 shows the computer simulated results when the cross talk effect is modeled by mutual inductors. As the error between the Chua’s circuits \( \alpha_{1k} - \alpha_{2k} \) increases, the solution escapes from the synchronized plane more often. Figure 5(c) shows the time waveform of \( x_{10} + x_{20} \). We can see from this figure that the solution behaves under the on-off intermittency.

Figure 6 shows the computer simulated results when the cross talk effect is modeled by coupling capacitors. We can observe the similar phenomena to the case of mutual inductors. However, the breakdown seems to be less sensitive to the error of the parameters than the case of mutual inductors.

4. Cross Talk from Opposite Direction

Now, we consider the two Chua’s circuits with lossless transmission lines placed in parallel but from the opposite direction as shown in Fig. 7. Similar to the circuit model in Fig. 2, we model the lossless transmission lines by LC ladder circuits with finite numbers of lumped elements as shown in Fig. 8.

Further, we consider that the effect of the cross talk is modeled by mutual inductors as shown in Fig. 9. We carry out computer simulation for this models by using the 4th order Runge-Kutta method. In this simulation, we consider only the case that the two Chua’s circuits are identical. So, we can rewrite the parameters as follows:

\[
\begin{align*}
\alpha_1 &= \alpha_2 = \alpha, \quad \gamma_1 = \gamma_2 = \gamma, \quad a_1 &= a_2 = a, \\
b_1 &= b_2 = b, \quad c_1 &= c_2 = c, \quad d_1 &= d_2 = d. 
\end{align*}
\]

In the following simulations, we fix the parameters as follows:

\[
\begin{align*}
a &= -1.2, \quad b = -0.75, \quad c = 10, \quad d = 8, \\
\alpha &= 18, \quad \gamma = 1, \quad n = 10. 
\end{align*}
\]

4.1. Cross Talk via Mutual Inductor \( M \)

Figure 10 shows the synchronization of the two circuits. It is interesting to observe that the two circuits can be synchronized even if they are placed from the opposite direction. The parameter \( m \) corresponds to the overlap of the two transmission lines. Namely, \( m = 10(= n) \) means that all of inductors are coupled to one of the other inductors via mutual inductors. While \( m = 5 \) means that only the half of the transmission lines is influenced each other. Further, \( m=1 \) means that only the last inductors in the two transmission lines are coupled each other.

4.2. Cross Talk via Mutual Inductor \( -M \)

Next, we consider the case that the cross talk effect is modeled by negative mutual inductors. Figure 11 shows

\[\text{Figure 5: Breakdown of synchronization. Cross talk effect is modeled by mutual inductors.}\]

\[\text{Figure 11: Breakdown of synchronization. Cross talk effect is modeled by negative mutual inductors.}\]
Figure 7: Two Chua’s circuits with transmission lines placed from the opposite direction.

Figure 8: Discrete model by LC ladder circuits with finite number of lumped elements.

Figure 6: Breakdown of synchronization. Cross talk effect is modeled by coupling capacitors.

the simulation results for the case that \( M \) is negative. In this case, we confirm that the two circuits synchronize in in-phase.

4.3. Effect of Coupling Strength

Finally, we carry out computer simulation as varying \( \beta \), namely, the coupling strength. Figure 12 shows the simulation result when \( m = 3 \). From Fig. 12, we can say that the synchronization is robust against the coupling strength.

While Fig. 13 shows the simulation result when \( m = 5 \). This results show that the synchronization is sensitive to the change of the coupling strength.

At the moment, we can not explain why the robustness of the synchronization against the coupling strength is changed according to the position of the coupling of the transmission line.

5. Conclusions

In this study, we have considered the two Chua’s circuits with lossless transmission lines placed in parallel. We have modeled the effect of the cross talk by mutual inductors. By computer simulations, we have investigated various interesting phenomena related with chaos synchronization.

In our future work, we investigate Chua’s circuit with lossy transmission line, because real transmission lines should have loss actually. Furthermore, we investigate crosstalk phenomena between conductor boards placed in parallel and apply the result to chaotic circuits.
Figure 9: Model of cross talk effect by mutual inductors.

Figure 11: Simulation results of cross talk via mutual inductor $-M$.

Figure 12: Effect of coupling strength when $m = 3$.

Figure 13: Effect of coupling strength when $m = 5$.

References


Dual Synchronization of Chaos in Mackey-Glass Electronic Circuits with Time Delay

Satoshi Sano, Atsushi Uchida, and Shigeru Yoshimori

†Department of Electronics and Computer Systems, Takushoku University
815-1 Tatemachi, Hachioji, Tokyo 193-0985, Japan
Email: sanosatoshi2005@yahoo.co.jp

Abstract— We experimentally and numerically demonstrate the dual synchronization of chaos in two pairs of one-way coupled Mackey-Glass electronic circuits. The output of the two drive circuits are mixed and used both for the feedback signal to the two drive circuits and for the transmission signal to the two response circuits. We investigate the regions for achieving dual synchronization of chaos when the delay time is mismatched between the drive and response circuits.

1. Introduction

Synchronization of chaos in electronic circuits have attracted increase interests for the applications of secure communications [1,2] and spread spectrum communications [3]. Synchronization of chaos can be used for sharing the identical chaos in a transmitter and a receiver as a cryptographical code. Many studies on synchronization of chaos have been reported in one-way coupled chaotic systems [1,2]. However, since the configuration of synchronization is limited to a single pair of one-way coupled oscillators, this method cannot be applied for multi-user communication systems [4].

The technique of multiplexing is a very important issue for high capacity communications [4]. Multiplexing chaos using synchronization has been reported in a simple map and electronic circuit model [5]. Dual synchronization of chaos for synchronizing two different pairs of chaotic maps and delay-differential equations has also been investigated [6]. To synchronize each pair of chaotic systems, all the parameter settings must be identical between the transmitter and the receiver, whereas they must be slightly shifted between different pairs of chaotic systems. We have recently reported dual synchronization of chaos in both microchip solid-state lasers and Colpitts electronic oscillators [7,8]. However, all the response systems need to be synchronized to achieve dual synchronization in electronic circuits [8].

In this paper, we experimentally demonstrate the dual synchronization of chaos in two pairs of one-way coupled Mackey-Glass electronic circuits with time delay. The degree of synchronization is quantitatively evaluated by using the cross correlation. We investigate the parameter regions for achieving dual synchronization when the delay time is mismatched between the drive and response circuits.
2. Experiments

2.1. Setup

Figure 1 shows the diagram of the Mackey-Glass electronic circuit. The circuit consists of two field effect transistors (2SJ103 and 2SK30A) for the nonlinear part, three amplifiers, a RC filter, and a delay line. Fifty pairs of capacitors (10 nF) and inductors (4.7 mH) are combined to generate time delay. The total delay time corresponds to a half period of chaotic waveforms.

Figure 2 shows a block diagram for dual synchronization of chaos in Mackey-Glass electronic circuits with time delay. Two of the electronic circuits are used as drive systems (Drive 1 and Drive 2). The other two circuits are used as response systems (Response 1 and Response 2). The voltages in two drive systems are added and fed back to each drive system with time delay to generate chaotic oscillations. This feedback signal is also used as a transmission signal, and the signal is transmitted to the two response systems for dual synchronization. There exists symmetry between the drive and response systems in terms of the input signal. All the parameters need to be matched between the Drive 1 and Response 1, and between the Drive 2 and Response 2, whereas the delay times for the Drive 1 and Drive 2 are set to be different for achieving dual synchronization (228 and 307 μs, respectively). When synchronization manifolds are stable between the Drive 1 and Response 1, and between the Drive 2 and Response 2, dual synchronization can be achieved.

Fig.4 Experimental results of temporal waveforms and their correlation plots. (a) (b) Drive 1 and Response 1, (c), (d) Drive 2 and Response 2, and (e), (f) Drive 1 and Response 2. Dual synchronization is achieved in the pair of Drive1-Response1 and Drive2-Response2.
2.2. Experimental results

Without coupling between the drive and response circuits, chaotic oscillations are observed through some bifurcation processes in all the circuits. Figure 3 shows chaos attractors generated in Drive 1 and Drive 2. Two different chaos attractors are observed for the two drive circuits with different delay times (228 and 307 µs, respectively). In the presence of coupling from the drive to response systems, dual synchronization is observed. Figure 4 shows temporal waveforms of the voltages for all the four electronic circuits and their correlation plots of the pairs of Drive1-Response1, Drive2-Response2, and Drive1-Response2. Synchronization of chaotic oscillations is independently achieved for the pairs of Drive1-Response1 and Drive2-Response2, as shown in Figs. 4(a) - 4(d). A good linear correlation is observed as shown in Figs. 4(b) and 4(d). Conversely, synchronization of chaos is not achieved for the pair of Drive1-Response2, as shown in Figs. 4(e) and 4(f).

2.3. Parameter dependence

We quantitatively investigate chaos-synchronization regions against parameter mismatch in the two pairs of Mackey-Glass electronic circuits. The delay times of Drive 1, Drive 2, and Response 1 are fixed at Drive 1 delay time of Drive1 327 µs, delay time of Response 2 327 µs, and, delay time of Drive 1 of 398 µs respectively. The delay time of Response 2 is shifted. Other parameters in the four circuits are set to be as identical as possible, although there still exist some errors in our experimental systems. Figure 5 shows the cross correlation of the two temporal waveforms for the pair of Drive1-Response2 and for the pair of Drive2-Response2. We found that synchronization can be achieved when the delay time for Response 2 is set to be equal to the delay time for either Drive 1 or Drive 2. The best correlation is obtained with the matched delay time for both of the pairs of Drive1-Response2 and Drive2-Response2. This demonstrates that mismatched delay times reduce the accuracy of synchronization. We also found that it is important to set parameters significantly different between Drive 1 and Drive 2 for the achievement of dual synchronization.

2.4. Simulation

To explain our experimental observation, we numerically calculate a model for Mackey-Glass electronic circuits. The model equation is as follows.

\[ \frac{dx(t)}{dt} = \frac{acx(t-\tau)}{1+cx^n(t-\tau)} - bx(t) \cdots \cdots \cdots \cdots \quad (1) \]

There are five parameters: \(a\) (height of the nonlinear function), \(b\) (RC constant), \(\tau\) (delay time), \(n\) (shape of the nonlinear function), and \(c\) (feedback strength). We set these parameters as follows: \(a=3.2, n=4.5, \tau=8.2, b=1.0, c=1.0\) for Drive 1 and Response1, and \(a=2.0, n=10.0, \tau=11.0, b=1.0, c=1.0\) for Drive 2 and Response 2. Figure 6 shows chaotic attractors generated in Drive 1 and 2. These attractors look similar to those observed in our experiment (see Fig. 3). Figure 7 shows numerical results of temporal waveforms and their correlation plots. Chaos synchronization is achieved between Drive 1 and Response1, and between Drive 2 and Response 2 as shown in Figs. 7(a)-7(d). Conversely, no synchronization is observed between Drive 1 and Response 2 (Figs. 7(e) and 7(f)).

Fig.5 Cross correlation as a function of the delay time for Response 2. Solid line: cross correlation between Drive 1 and Response 2. Dotted line: cross correlation between Drive 2 and Response 2. The vertical arrows indicate the parameter matching condition of the delay time.

Fig.6 (a) Chaos attractor of Drive1 generated by simulation. (b) Chaos attractor of Drive2 generated by simulation
3. Conclusion

We have experimentally and numerically demonstrated the dual synchronization of chaos in two pairs of one-way coupled Mackey-Glass electronic circuits. The output of the two drive circuits are mixed and used both for the feedback signal to the two drive circuits and for the transmission signal to the two response circuits. We have investigated the regions for achieving dual synchronization of chaos when the delay time is mismatched between the drive and response circuits. The technique of dual synchronization could be useful for multiplexing communications using chaos and for the identification of chaos from mixed chaotic waveforms.

Fig. 7 Numerical results of temporal waveforms and their correlation plots. (a) (b) Drive 1 and Response 1, (c), (d) Drive 2 and Response 2, and (e), (f) Drive 1 and Response 2. Dual synchronization is achieved in the pair of Drive1-Response1 and Drive2-Response2.

References


A Master Stability Function Approach for Separation and Synchronization of Chaotic Systems

Arturo Buscarino, Luigi Fortuna and Mattia Frasca

Università degli Studi di Catania, viale A. Doria 6, 95125 Catania, Italy.
Email: lfortuna@diees.unict.it

Abstract—In this paper, by using a negative feedback scheme we study the problem of synchronizing two systems, each of them made of \( n \) independent piecewise linear (PWL) chaotic circuits, through the transmission of a unique signal. The signal is the linear combination of the chaotic signals generated by the \( n \) circuits and, thus, to synchronize the two groups of chaotic circuits the slave circuit has first to separate the contributions of the subunits constituting the master system. To find the appropriate values of the feedback gains, an approach based on the master stability function and an optimization procedure through genetic algorithms is introduced. Numerical results showing that with this approach separation and synchronization of \( n = 3 \) chaotic systems can be achieved are discussed.

1. Introduction

Since the seminal work by Pecora and Carroll [1] on synchronization of chaotic systems, many schemes for synchronization have been proposed [2]: negative feedback [3], sporadic driving, active-passive decomposition, diffusive coupling and some other hybrid methods. Moreover, both bidirectional and unidirectional couplings have been considered [2]. In the case of unidirectional coupling, the evolution of one of the coupled systems is unaltered by the other system. In this case, the two chaotic systems are called master and slave or drive and response systems.

In the negative feedback scheme [3], an error signal, built as the difference of two corresponding state variables (which are assumed measurable), is fed back into the slave system. In this paper the simultaneous synchronization of two groups of \( n \) chaotic subsystems instead of two chaotic systems is investigated by using a negative feedback scheme. In our case, thus, the master system is formed by \( n \) independent chaotic systems which do not interact each other. In general, the synchronization of two groups of such chaotic systems requires \( n \) independent feedback signals. In this paper it is investigated if and under which conditions synchronization can be achieved by using a unique feedback signal which depends on the \( n \) chaotic systems of the master (i.e. it is for instance a linear combination of the state variables of the master chaotic systems). This problem is referred to as separation and synchronization of chaotic signals.

A theoretical approach, based on the investigation of the synchronization properties through master stability function (MSF) analysis [4] of multiplexed systems, is discussed. The synchronization properties of two multiplexed systems depend on the choice of the variables used to build the scalar signal. Given a fixed choice of these variables, the analysis based on the MSF allows us to investigate if synchronization can be achieved. When such synchronization cannot be achieved, the choice of the parameters (state variables and weights of the linear combination) can be optimized. The strategy proposed in this paper is based on genetic algorithms with a fitness function based on the MSF of the considered system.

Some examples for the case of \( n = 2 \) are shown. For the case of \( n = 3 \) a numerical simulation showing the separation of the three chaotic dynamics and the synchronization of the three pairs of chaotic systems is reported.

A similar topic, referred as multiplexing of chaotic signals using synchronization, was investigated in [5] and [6]. In particular, Tsimring and Sushchik [6] investigate the simultaneous synchronization of chaotic maps, while Carroll and Pecora discuss how to combine two chaotic systems with the multiplexing technique to make a communication system.

Furthermore, the solution to the problem of separation and synchronization, proposed in this paper, can be adopted in chaotic communication systems.

2. Master stability function for multiplexed systems

In this paper we investigate the synchronization properties of systems made of \( n \) different and uncoupled chaotic subunits. These systems are coupled through scalar signals obtained by summing the state variables of the chaotic subunits. The scalar signal can then be transmitted into a unique channel used to synchronize multiple independent chaotic subsystems. Since multiple signals are sent to the same channel, we refer to these systems as multiplexed systems.

We first describe how the scalar signal is used to connect two multiplexed systems. The scheme proposed is based on negative feedback. Let us consider two multiplexed systems as shown in Fig. 1, where a directional coupling from the multiplexed system A to the multiplexed system B is taken into account. The two multiplexed systems are considered identical, i.e. they contain the same \( n \) independent
Subsystems with equal parameters. A linear combination of the state variables of system A is sent to system B, where an error signal is built by comparing the received signal with the same linear combination of the corresponding state variables of system B.

The error signal is weighted by suitable gains and added to each state variable of system B as in the negative feedback scheme for two chaotic systems [3].

Therefore, assuming that the equations of system A are:

$$\dot{X}_A = f(X_A),$$

system B will be described by the following equations:

$$\dot{X}_B = f(X_B) + K e,$$

where $K$ is the gains vector and $e$ is the (scalar) error signal. Assuming that each multiplexed system is composed by $n$ systems of order $m_1, m_2, \ldots, m_n$, then $X_A \in \mathbb{R}^m$ with $m = m_1 + m_2 + \ldots + m_n$, $X_B \in \mathbb{R}^m$ and $K \in \mathbb{R}^m$.

The synchronization properties of multiplexed systems can be studied with the master stability function (MSF). The approach, introduced in [4], considers $N$ identical oscillators coupled with the same function of the components from each oscillator to the other oscillators into an arbitrary network which admits the synchronization manifold as an invariant manifold. The approach is based on the linearization of the network dynamics around the synchronization manifold.

In [4] the dynamics of each node is modelled as $\dot{x}^i = F(x^i) + \alpha \sum_j G_{ij} H(x^j)$ where $\dot{x}^i = F(x^i)$ represents the dynamics of each isolated node, $\alpha$ is the coupling strength, $H : \mathbb{R}^m \to \mathbb{R}^m$ the coupling function and $G = [G_{ij}]$ is a zero-row sum matrix modelling network connections. The synchronization properties of this network are studied by calculating the maximum Lyapunov exponent $\lambda_{max}$ of the generic variational equation

$$\dot{\xi} = [DH + (\alpha + i\beta)DH] \xi$$

as a function of $\alpha$ and $\beta$, where $DF$ and $DH$ represent the Jacobian of $F(x')$ and $H(x')$ computed around the synchronous state. Once obtained $\lambda_{max}$, which does not depend on the connection network, the stability of the synchronization manifold in a given network can be evaluated by computing the eigenvalues $\gamma_h$ (with $h = 2, \ldots, N$) of the matrix $G$ and studying the sign of $\lambda_{max}$ at the points $\sigma \gamma_h$. If all eigenvalues with $h = 2, \ldots, N$ are stable, then the synchronous state is stable at the given coupling strength. In fact, we recall that, since $G$ is zero-row sum, the first eigenvalue is $\gamma_1 = 0$ and represents the variational equation of the synchronization manifold.

In particular, if $G$ has real eigenvalues, as in our case, the master stability function can be computed only as function of $\alpha$. In the following we will restrict our analysis to this case. The functional dependence of $\lambda_{max}$ on $\alpha$ can give rise to three different cases [7]. The first case, denominated as type I, is the case in which network nodes cannot be synchronized. In the second case (type II) increasing the coupling coefficient $\sigma$ always leads to a stable synchronous state. In the third case (type III), network nodes can be synchronized only if $\sigma \gamma_h$ for $h = 2, \ldots, N$ lie in the interval with negative values of $\lambda_{max}$.

Referring to the formulation of the synchronization problem of multiplexed systems in terms of equations (2), the matrix $DH$ becomes

$$DH = \begin{bmatrix} k_1 b_1 & k_1 b_2 & \ldots & k_1 b_m \\ k_2 b_1 & k_2 b_2 & \ldots & k_2 b_m \\ \vdots & \vdots & \ddots & \vdots \\ k_m b_1 & k_m b_2 & \ldots & k_m b_m \end{bmatrix}$$

where $k_1, k_2, \ldots, k_m$ are the gains ($K = \{ k_1, k_2, \ldots, k_m \}$) and $b_1, b_2, \ldots, b_m$ with $b_i = \{ 0, 1 \}$ specify which state variables are used to build the error signal $e$. Let us define $B = \{ b_1, b_2, \ldots, b_m \}$.

The master stability equation will in general depend on these parameters $K$ and $B$. At this point, the problem of the possibility of synchronizing multiplexed systems can be translated into the problem of the existence of suitable values of $K$ and $B$ for which the master stability function is either type II or type III.

To solve this problem, an approach based on genetic algorithms can be used. Genetic algorithms are an optimization procedure based on the evolution of a population of individuals coding the possible solutions to the problem [8]. During the population evolution, crossover and mutation ensure the search for new solutions, while selection is performed according to a given fitness function which defines the optimization problem. In our case the problem is to obtain a master stability function with $\lambda_{max} < 0$. To this aim the fitness function can be defined as follows:

$$f = \min \lambda_{max}$$
Given this fitness function, the genetic algorithms are used to search \( K \) and \( B \) which minimize the maximum Lyapunov exponent of Eq. (3). If the optimum value is such that \( \lambda_{\text{max}} < 0 \), then the problem of synchronizing multiplexed systems will have a solution. Of course, the existence of this solution is not related to the stability of the synchronization manifold in a given complex network, but \( \lambda_{\text{max}} < 0 \) ensures that there exists synchronizable networks with multiplexed systems as nodes. For instance, if two multiplexed systems are coupled in a master-slave configuration (i.e. \( G = \begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \) and \( \gamma_1 = 0, \gamma_2 = -1 \)), \( \lambda_{\text{max}} < 0 \) ensures that there exists a suitable value of the coupling strength \( \sigma \) such that \( \lambda(\sigma) < 0 \) with \( \sigma = \sigma \gamma_2 = -\sigma \) for which the two systems will synchronize.

3. Numerical results

The proposed approach has been applied to multiplexed systems of three well-known chaotic systems: Lorenz system, Rössler oscillator and Chua’s circuit. The dimensionless equations describing these three systems are reported in the following with the parameters adopted in the paper. All the parameters are chosen so that chaotic behavior is obtained.

The Lorenz system is described by the following equations:

\[
\dot{x} = \sigma(y - x);
\dot{y} = px - y - xz;
\dot{z} = xy - rz
\]

with \( \sigma = 10, p = 8/3, r = 28 \).

The following dimensionless equations model the Rössler oscillator:

\[
\dot{x} = -(y + z);
\dot{y} = x + ay;
\dot{z} = b - x + cy
\]

The Chua’s circuit is described by the following equations:

\[
\dot{x} = \sigma(y - x - z);
\dot{y} = x - cy - z
\]

\( \sigma = 10, b = 16, a = 1, c = -0.143 \).

We applied the MSF approach to several multiplexed systems obtained by pairing two of the three chaotic systems described above. Moreover, these multiplexed systems differ from the way in which the two subsystems are coupled. In other words depending on the definition of the error scalar signal adopted, different multiplexed systems are obtained, which in general have different synchronization properties.

In the following, the notation adopted to define the error signal will be \( e = E(x_1, \ldots, x_n, y_1, \ldots, y_n; z_1, \ldots, z_n) \) where

\[
E(x_1, \ldots, x_n, y_1, \ldots, y_n, z_1, \ldots, z_n) = \\
x_{1,m} - x_{1,s} + \ldots + x_{n,m} - x_{n,s} + \\
y_{1,m} - y_{1,s} + \ldots + y_{n,m} - y_{n,s} + \\
z_{1,m} - z_{1,s} + \ldots + z_{n,m} - z_{n,s}
\]

and \( x_1, \ldots, x_n, y_1, \ldots, y_n, z_1, \ldots, z_n \) represent all the subsystem state variables effectively used to build the error signal. For instance, if we consider two subsystems (a Lorenz system and a Chua’s circuit), coupled through \( x_L \) and \( x_C \), thus \( E(x_1, \ldots, x_n, y_1, \ldots, y_n, z_1, \ldots, z_n) = E(x_L, x_C) \) where:

\[
E(x_L, x_C) = x_{L,m} - x_{L,s} + x_{C,m} - x_{C,s}
\]

Table 1: Synchronization properties of multiplexed systems made of two chaotic subsystems with respect to different definition of the error signal with \( K = K_1 \).

| & \( E_{xx} \) & \( E_{xy} \) & \( E_{xz} \) & \( E_{xxy} \) & \( E_{xzx} \) & \( E_{yxy} \) & \( E_{yzx} \) |
|---|---|---|---|---|---|---|---|
| L-C & I & I | III | I | I | I | I |
| C-R & I & I | I | III | I | I | I |
| L-R & I | III | I | I | I | I | III |

The first case analyzed was the case in which we fixed \( K = K_1 = \left[ \begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{array} \right] \). Based on the results obtained, the multiplexed systems can be classified according to their synchronization properties reported in Table 1.

As an example, we can consider the case of a multiplexed system made of a Lorenz system and a Rössler oscillator, coupled through \( E(y_L, y_R) \) which as shown in Table 1 is type III.

Considering only matrices \( G \) with real eigenvalues, the MSF behavior for such multiplexed system is type III. Thus, exists a range of \( \alpha \) corresponding to negative values of \( \lambda_{\text{max}} \). This means that synchronization can be achieved choosing the coupling strength \( \sigma \) in order to put all the eigenvalues \( \sigma \gamma_k \) of matrix \( G \) inside such range.

As it can be noticed in Table 1, in most cases the multiplexed system is type I and therefore synchronization cannot be achieved. This can be due to the choice of the coupling parameters (i.e. \( K \)). For this reason, when the subsystem is type I, we investigated the possibility of finding a coupling vector such that the multiplexed system is type II or III by applying genetic algorithms with the fitness function defined as in (5). With this approach it is effectively possible to find suitable parameters in several cases.

For instance, in the case of a multiplexed system made of a Lorenz system and a Rössler oscillator, coupled through \( E(x_L, y_L, y_R, x_R) \) (which is type I for \( K = K_1 \)), we were able to find a suitable coupling vector \( K = K_{LR} \) such that the multiplexed system is type III. Genetic algorithms (GA) were used with cross-over probability \( p_c = 0.7 \), mutation probability \( p_m = 0.7 \) and generation gap \( g_{gap} = 0.9 \). The elements of vector \( K \) were searched in the range \([-2; 2]\) generating 20 individuals that evolve for 30 generations. Chromosomes are represented with 16 bit precision, thus it is possible to select \( 2^{16} \) values inside the fixed range.

The GA minimizes the fitness function (5) finding the vector \( K_{LR} = \left[ \begin{array}{c} 1.43 \\ 1.25 \\ -0.13 \\ 0.96 \\ 0.10 \end{array} \right] \) the resulting MSF is type III. This means that using \( K = K_{LR} \) the considered multiplexed system can be synchronized setting \( \sigma \) such that the eigenvalues \( \sigma \gamma_k \) of \( G \) lie inside the range of \( \alpha \) corresponding to the negative values of the MSF.

Using GAs, we were also able to obtain a set of gains
4. Conclusions

In this paper, the problem of separation and synchronization of chaotic systems has been investigated by using an approach based on the master stability function. With this approach combined with an optimization strategy based on genetic algorithms, it is possible to investigate if two multiplexed chaotic systems in a master-slave configuration can be synchronized and to find the suitable parameters leading to synchronization. Furthermore, the approach is more general and allows one to draw conclusion on the synchronization properties of networks in which each node is given by a system made of $n$ independent chaotic subunits.

References

SYNCHRONOUS REGIMES IN A CHAIN OF COUPLED BONHOEFFER-VAN DER POL OSCILLATORS

G.V.Osipov, A.K.Kryukov, A.V.Polovinkin† and J.Kurths‡

†Radiophysics Dept., University of Nizhny Novgorod Gagarin ave. 23, Nizhny Novgorod, Russia, 603950
‡Institute of Physics, University Potsdam, 10, Am Neuen Palais, D-14415, Potsdam, Germany
Email: osipov@rf.unn.ru, polovinkin@rf.unn.ru, juergen@agnd.uni-potsdam.de

Abstract—We study synchronous behavior in ensembles of locally coupled non-identical Bonhoeffer - van der Pol oscillators. We show that in a chain of $N$ elements $2^{N-1}$ different regimes of global synchronization are possible.

1. Introduction

The understanding of principles of functioning of the human neuronal system and algorithms of information processing in neuron systems is an important actual challenge. Answers to these problems will have an immediate impact on the creation of highly efficient and low cost artificial neuron systems which are capable to solve tasks, apparent now as extremely complex [1, 4]. There are already first solutions in this direction demonstrating the potentials of artificial networks constructed by analogy with neuron systems. For example, it is processing of threads of multimedia data, including tasks of recognition of texts and images, optimum management of complex structures, brain-machine interactions etc.

One basic feature of ensembles of neurons in the central and peripheral nervous systems, or in cardiac tissue is there ability to synchronization [3, 4, 5, 6, 7, 8]. Therefore, the study of synchronization in chains and networks of elements simulating self-oscillatory activity of neurons and cardiac cells is extremely important. In this paper we investigate synchronization in small (two and three elements) and large (chain) ensembles of coupled neuron-like oscillators. We demonstrate that such ensembles generate multistability of synchronous regimes. In dependence on the initial conditions in a chain of $N$ coupled oscillators, $2^{N-1}$ synchronous regimes are possible. For $N = 2$ the existence of in-phase and anti-phase regimes is proved analytically. Numerical simulations show the appearance of four different synchronous regimes for $N = 3$. In large ensembles different regimes of global and cluster synchronization are found.

2. Model

In this paper we investigate a chain of locally diffusively coupled Bonhoeffer - van der Pol (BvdP) oscillators as a model of a neuron network [8]:

$$\begin{cases}
\dot{x}_j = F_j(x_j, y_j) + d(x_{j+1} - 2x_j + x_{j-1}), \\
\dot{y}_j = G_j(x_j),
\end{cases} \quad j = 1, ..., N,$$

where $F_j(x_j, y_j) = x_j - 1/3 x_j^3 - y_j$, $G_j(x_j) = \epsilon (x_j - a_j)$, $N$ is the number of elements in the chain, and $d$ is the coupling between the elements, $\epsilon < < 1$, $0 < a_j < 1$. We consider slightly non-identical oscillators, which have small parameter mismatches $\Delta a_j = a_j - a_j$. Free-end boundary conditions are taken. Because $\epsilon < < 1$ is very small, all oscillations in (1) can be divided in slow and fast motions. From a physiological point of view, the fast variable $y$ can be considered as a voltage, but the slow variable $x$ - as a gating or recovery variable. Therefore, the cooperative behavior of neuronal nets or cardiac tissue can be at least qualitatively reproduced by a model of coupled BvdP oscillators. In [9] this oscillator were successfully used as individual cell for modeling the frequency entrainment of heart pacemakers. Collective dynamics of synaptically coupled BvdP neurons was studied in [10].

3. Single neuron

For a single element:

$$\begin{cases}
\dot{x} = x - y - x^3/\tau, \\
\dot{y} = \epsilon (x + a),
\end{cases}$$

there is one unstable steady state $(\bar{x}, \bar{y}) = (-a, a^3/\tau - a)$. It is (i) a focus if $a > \sqrt[3]{1 - 2 \sqrt{6}}$, and (ii) a node if $a < \sqrt[3]{1 - 2 \sqrt{6}}$. There exists also a stable limit cycle. Because of the smallness of $\epsilon$ it consists of fast and slow parts. A typical phase portrait is shown in a Fig.1a. Here $h_-(x)$ and $h_+(x)$ are the left resp. right steady parts and $h_0(x)$ is the unstable part of the curve of slow movements, correspondingly.

4. Two coupled neurons

For some fixed parameters there are two limit cycles in a pair of coupled elements [6,7]. These cycles correspond to
in-phase and anti-phase synchronous regimes. For the analytical proof of this fact for $a_1 \approx a_2$, we consider a piecewise linear approximation of the functions $F_1, 2$ (Fig. 1b):

$$ F_i = \begin{cases} 
-4 \cdot \frac{x_i - 2 - y_i}{3}, & \text{for } x_i \leq -1, \\
2 \cdot \frac{x_i - y_i}{3}, & \text{for } -1 < x_i < 1, \\
-4 \cdot \frac{x_i + 2 - y_i}{3}, & \text{for } x_i \geq 1.
\end{cases} \tag{3} $$

Solving this linear system, we find both limit cycles. To do this, we build the mapping $\frac{df_p}{df_q}$, while $n$ corresponds to the $n$th passing of the limit cycle, and $df_p = y_1^n - y_2^n$. Without losing generality, let us consider the following case. Let $y_1^n, y_2^n$ be the values of $y_1$ and $y_2$ on the $n$th passing of the limit cycle. Then the first element just comes to the line $h_+(x)$. The state $y_2^n$ of the second element can be arbitrary on the lines $h_-(x)$ or $h_+(x)$. The state $y_1^n$ of the first element is defined through $y_2^n$. Let us assume that the second element is located on the $h_+(x)$. Then

$$ y_1^n = 2/3 + d(x_2 - x_1) = \hat{d}^{-1}(4\hat{d}/3 - dy_2^n), \tag{5} $$

where $\hat{d} = 2/3 + d$ and $\hat{d} = 4/3 + d$. Let in the moment $\tau = t_1$ the second element jumps to the line $h_-(x)$. Then, solving system (4), one obtains for $y_1(t_1)$ and $y_2(t_1)$:

$$ y_1^{-}(t_1) = \left[ \frac{1}{2}(y_1^n - y_2^n) + 2 - (a_1 - a_2)d \right] \exp(-\frac{\hat{d}}{2\hat{d}}) + \left[ \frac{1}{2}(y_1^n + y_2^n) - \frac{3}{2}(a_1 + a_2) \right] \exp(-\frac{\hat{d}}{4\hat{d}}) - 2 - da_2 + \hat{d}a_1, \tag{6} $$

$$ y_2^{-}(t_1) = \left[ \frac{1}{2}(y_1^n - y_2^n) - 2 + (a_1 - a_2)d \right] \exp(-\frac{\hat{d}}{2\hat{d}}) + \left[ \frac{1}{2}(y_1^n + y_2^n) - \frac{3}{2}(a_1 + a_2) \right] \exp(-\frac{\hat{d}}{4\hat{d}}) + 2 - da_2 + \hat{d}a_1. \tag{7} $$

In expressions (6) and (7) the index "−" means, that the first element is on the line $h_+(x)$, and the second one on the line $h_-(x)$. From another side, because the fact that at $\tau = t_1$ the second element jumps to the line $h_-(x)$, we have:

$$ y_2(t_1) = \frac{2}{3} + \frac{d}{2\hat{d}}(y_2(t_1) - y_1(t_1) - 4) \tag{8} $$

Comparing (7) and (8) and taking into account (6), we can find the moment $t_1$. Now both elements are on the line $h_-(x)$. The corresponding values $y_1(\tau)$ and $y_2(\tau)$ can be calculated:

$$ y_1^{-}(\tau) = \left[ \frac{1}{2}(y_1(\tau_0) - y_2(\tau_0)) - (a_1 - a_2)d \right] \exp(-\frac{\hat{d}}{2\hat{d}})(\tau - \tau_0) + \left[ \frac{1}{2}(y_1(\tau_0) + y_2(\tau_0)) + 2 - \frac{3}{2}(a_1 + a_2) \right] \exp(-\frac{\hat{d}}{4\hat{d}})(\tau - \tau_0) - 2 - da_2 + \hat{d}a_1, \tag{9} $$

$$ y_2^{-}(\tau) = \left[ \frac{1}{2}(y_2(\tau_0) - y_1(\tau_0)) + (a_1 - a_2)d \right] \exp(-\frac{\hat{d}}{2\hat{d}})(\tau - \tau_0) + \left[ \frac{1}{2}(y_1(\tau_0) + y_2(\tau_0)) - 2 - \frac{3}{2}(a_1 + a_2) \right] \exp(-\frac{\hat{d}}{4\hat{d}})(\tau - \tau_0) - 2 - da_2 + \hat{d}a_1. \tag{10} $$

Figure 1: a). Phase portrait of system (2). b). In-phase regime at piece-wise linear approximation of the functions $F_1, 2$ given according to (3).
Each element can reach the left extreme before the other element. This depends on the initial conditions and the parameters. Let at \( \tau = t_2 \) one of the elements reaches the left extreme. Then \( y_1(t_2), y_2(t_2) \) can be obtained from (9) and (10). From another side, we can use the fact that one element at \( \tau = t_2 \) is located in the left extreme. Therefore, we can define the moment \( t_2 \). If we move further along the cycle, then we can obtain (i) the moment where the first element jumps again from the line \( h_+(x) \) to the line \( h_-(x) \), and (ii) the values \( y_1^{\text{pred}} \) and \( y_2^{\text{pred}} \). Therefore, we can build the mapping \( df_{\text{fin}}(df_a) \). This mapping at \( d = 0.002 \), \( a_1 = 0.995 \), \( a_2 = 0.994 \) and at the initial state of the second element the \( h_-(x) \) is presented in Fig. 2a. It is visible, that there are two steady fixed points, each related to the certain synchronising regime. The fixed point in a vicinity of zero corresponds to an in-phase regime \( (x_1 \approx x_2) \); the point near \( df(n) = 1.323 \) corresponds to an anti-phase regime. Both of these regimes appear if the coupling strength \( d \) becomes larger than some critical value. With further increase of \( d \), the fixed point corresponding to an anti-phase regime disappears but the fixed point corresponding to an in-phase regime remains (Fig. 2b). This means that for relatively large coupling only an in-phase synchronous regime exists.

Figure 2: Map \( df_{\text{fin}}(df_a) \). Initially the second element is located on \( h_+(x) \). Parameters are: \( a_1 = 0.995 \), \( a_2 = 0.994 \), \( \epsilon = 0.02 \) and: a) \( d = 0.002 \), b) \( d = 0.05 \).

These analytical results obtained for a linear approximation of the functions \( F_{1,2} \) are tested in numerical experiment with model (1) for \( N = 2 \), \( \epsilon = 0.02 \). This way we get the existence of in-phase and anti-phase synchronous regimes. The appropriate time series are given in Fig. 3a and 3b. It is necessary to note, that the anti-phase regime is realized not only in some intervals of the coupling parameter \( d \), but in some interval of the difference of the parameters \( a_2 - a_1 \). In the case of a large difference \( a_2 - a_1 \), when even for one of the elements the time of movement along \( h_+(x) \) is close to the time of movement along \( h_-(x) \), the anti-phase regime disappears. Therefore, it is possible to assume that the strong difference between the times of two parts of slow movement is a reason for the existence of two (and very big for large ensembles) synchronous regimes. We have obtained the evolution of the observed frequencies \( \omega_j = 2\pi/T_j \) vs. the parameter \( d \). Here \( T_j = 1/M \sum^{M}_{i=1} \tau_{ij} \)

where \( M \rightarrow \infty \), \( j = 1, 2 \), and \( T_{ij} \) is the sequence of time intervals between consecutive maxima of realization \( x_j(t) \). In other words \( \omega_j \) is an averaged frequency of the occurrence of maxima in the time series \( x_j(t) \).

As our numerical experiments show, the in-phase and anti-phase regimes have strongly different observed frequencies. The frequency of the in-phase regime is close to the maximal of the individual frequencies of the uncoupled elements, but the frequency of the anti-phase regime goes to zero if \( d \) increases, and therefore, the anti-phase regime disappears. It can be noticed, that for identical elements decreasing of this frequency can be estimated analytically. Thus, at \( d > d_{cr} \) only the in-phase regime exists.

5. Three coupled elements

In an ensemble of three coupled elements the interval of coupling, in which four synchronous regimes can be realized in dependence on the initial conditions, is found: (i) in-phase regime \((x_1 \approx x_2 \approx x_3, y_1 \approx y_2 \approx y_3)\) and three mixed regimes where two of the oscillators are synchronised in-phase

(ii) regime, for which \( x_1 \approx x_2, y_1 \approx y_2 \);

(iii) regime, for which \( x_2 \approx x_3, y_2 \approx y_3 \);

(iv) regime, for which \( x_1 \approx x_3, y_1 \approx y_3 \), in some sense this regime can be called anti-phase regime.

Therefore \( 2^{N-1} = 2^{3} = 4 \) synchronous regimes are possible. The lags between the time series \( x_1(t), x_2(t), x_3(t) \) are non constant and can be changed in dependence on the parameters. Under some conditions the regime of splay-state occurs, for which the time between the maxima of the time series is close to \( T/3 \), where \( T \) is the period of the synchronous oscillations.

The mixed regimes are established at some frequency \( \omega_{syst} \), such that \( \omega_{syst} < \omega_{max} < \omega_{in} \), where \( \omega_{syst}, \omega_{in} \) are the synchronization frequencies of the splay-state and the in-phase synchronous regimes, correspondingly. Mixed regimes exist in a wider interval of \( d \) than the splay state. With increasing \( d \) the mixed regimes disappear, and only the in-phase regime remains.
6. Global and cluster synchronization in large ensembles

In order to study synchronization phenomena in large ensembles of neuron-like elements, we investigate a chain of 50 coupled oscillators with linearly distributed parameters $a_j (a_j = a_1 + \Delta a (j-1))$. The results obtained in the previous sections allow us to suppose that in this chain for some parameters $2^{N-1}$ different regimes of global synchronization are possible. In numerical experiments several of such regimes were found. The evolution of the synchronization frequencies for an increasing coupling parameter is similar to the evolution in the systems of two and three coupled elements. As in the previous numerical experiments, in the in-phase regime the synchronization frequency is close to the maximal of the individual frequencies. This regime remains with increasing $d$. With an increase of coupling the formation of groups of synchronized neighboring elements, i.e. clusters of synchronization, appears. The number of clusters decreases with increasing $d$, and for $d > 0.008$ global synchronization sets in (Fig.5). The formation of synchronous clusters is observed for randomly distributed parameters $a_j$ as well.

![Figure 4: Global ($d = 0.008$) and cluster ($d = 0.001, 0.0035$) synchronization in the chain of 20 coupled elements for $a_1 = 0.995$, $\Delta a = 0.001$, $\varepsilon = 0.02$.](image)

7. Conclusions

Basing on the received results it is possible to assume, that in a system of $N$ locally diffusively coupled Bonhoeffer - Van der Pol oscillators for fixed values of parameters the number of different globally synchronous regimes can be equal to $2^{N-1}$. This was numerically confirmed for $N=3$. An analytical proof for the existence of two synchronous regimes was performed for $N=2$. In large ensembles a transition to global synchronization is accomplished with the formation of synchronization clusters. For relatively strong coupling, only the in-phase synchronous regime exists, which is realized on a frequency close to the maximal of the individual frequencies.

Many theoretical and experimental results show that synchronization phenomena play a very important role in brain activity. It is assumed that synchronous firing of neurons are an essential mechanism for information processing. Therefore, the observed multi-stability of synchronous regimes may be useful by understanding of mechanisms of different brain functions including image storage and recognition, visual perception, memory processing, control of movement and posture.

Acknowledgments

This work is done at financial support RFBR-NSC, project 05-02-90567 and RFBR-MF project 05-02-19815 and RFBR project 06-02-16996. G.O. and J.K. acknowledges support of EU-Network BioSim, Contract No. LSHB-CT-2004-005137.

References

Abstract—Hybrid systems are subject of recent interest for many applications. These systems are characterized by interactions of continuous and discrete dynamics, which give rise to complex behavior. Various models for specific subclasses of hybrid systems, such as hybrid automata, hybrid petri nets or switched systems, have been suggested so far. This paper summarizes a class of hybrid systems, namely the class of discretely controlled continuous systems, and proposes a general model for this class from which an embedded map and a return map is derived. We summarize different types of analysis that can be done by use of the return map. The paper has tutorial character.

1. Introduction

In this paper a class of hybrid systems, called discretely controlled continuous systems (DCCS), is investigated. Such systems are composed of a continuous and a discrete subsystem arranged in a feedback loop (Figure 1). The state of the discrete subsystem corresponds to the mode of operation of the overall system and provides the only controllable input to the continuous subsystem. Switching the mode changes the continuous dynamics. In general, the continuous subsystem possesses different equilibria for different discrete states. An event generator defines the interface between the continuous and the discrete-event dynamics. Depending on the value of continuous signals, such as time and the continuous state, this device generates events, which trigger discrete state transitions.

The presented structure is found in many applications, e.g. in switching power converters, combustion engines and anti-lock brakes.

The primary control objective in the design of a DCCS is usually specified in terms of an invariant region of operation $\mathcal{R}_o$, inside which the continuous state has to remain despite acting disturbances and parameter variations. Generally this region $\mathcal{R}_o$ does not contain any equilibria of the continuous dynamics. Hence, in each mode the continuous state is driven out of $\mathcal{R}_o$, which necessitates that discrete state transitions occur perpetually. Accordingly DCCS usually exhibit either periodic or chaotic behaviour. The control objective must be achieved by adequately switching among the modes of operation at the correct time instants. Additional implementational requirements limit the switching frequency, i.e. guaranteeing a minimal time span between switching events.

In this paper a hybrid model for DCCS is proposed. From this model a return map is derived, which samples the system’s behaviour at specific switching instants. Typical analysis and control problems for DCCS addressed in recent publications are listed, which can be investigated using the presented model.

2. Discretely controlled continuous systems

2.1. Hybrid model

A suitable hybrid model, which reflects the components of a DCCS is shown in Figure 1. The continuous subsystem is described by the equations

$$\dot{x}(t) = f(x(t), d(t), q(t)), \quad x(0) = x_0 \quad (1)$$

$$y(t) = h(x(t), d(t), q(t)) \quad (2)$$

where $x(t) : \mathbb{R} \to \mathbb{R}^n$, $y(t) : \mathbb{R} \to \mathbb{R}^m$, $d(t) : \mathbb{R} \to \mathbb{R}^n$ and $q(t) : \mathbb{R} \to \mathbb{Z}$ denote the continuous state and output signals, the disturbance input and the state of the discrete subsystem respectively. Function $f$ defines a Lipschitz-continuous vector field and $h$ is the output equation of the continuous process. At the moment of a discrete state transition $x(t)$ is discontinuous, but not $x(t)$. The discrete subsystem is described by the transition function $g$, which maps the current discrete state $\bar{q}(k)$ and the generated event $\bar{e}(k+1)$ to a successor state

$$\bar{q}(k+1) = g(\bar{e}(k+1), \bar{q}(k)), \quad \bar{q}(0) = q_0 \quad (3)$$

$$\bar{q}(t) = \bar{g}(k), \quad \text{for } t \in [\bar{t}(k), \bar{t}(k+1)). \quad (4)$$

Symbols with a bar always denote signal values at switching instants, which are enumerated by the integer $k \in \mathbb{N}$.
Each value of the piecewise constant signal \( q(t) \) corresponds to a mode of operation of the overall system. The event generator is given by an event function
\[
\Phi(x(t), q(t), t) = \min_{e \in E} \{ \phi(x(t), q(t), t, e) \} \leq 0
\] (5)
which selects the event \( e \in E \) from the set of possible events such that \( \phi(x(t), q(t), t, e) \) is minimal. Hence, the value of the piecewise constant signal \( e(t) \) is determined by
\[
e(t) = \arg \min_{e \in E} \phi(x(t), q(t), t, e).
\] (6)
At the time instant
\[
\ell(k+1) = \Phi_{\ell}^{-1}(x(t), q(t), t)|_{e=\ell(k+1)}
\] (7)
with \( T = \{ t \in \mathbb{R} : t > \ell(k) \land \Phi(x(t), q(t), t) = 0 \} \) \( \forall k \)
the corresponding event
\[
\tilde{e}(k+1) = \Phi_{\tilde{e}}^{-1}(x(t), q(t), t)|_{e=\tilde{e}(k+1)}
\] (9)
is issued by the event generator. Depending on the implemented control strategy, the event function (5) may depend on a subset of its arguments [1].

The model presented above can be cast into the framework of the well known hybrid automaton \( H = (Q, X, f, \text{Init}, I, E, G, R) \) [2], with
\[
Q \quad \text{finite set of discrete variables}
\]
\[
X \quad \text{finite set of } n \text{ continuous variables}
\]
\[
f : Q \times \mathbb{R}^n \to \mathbb{R}^n \quad \text{vector field}
\]
\[
\text{Init} \subseteq Q \times \mathbb{R}^n \quad \text{set of initial states}
\]
\[
I : Q \rightarrow I \subseteq \mathbb{R}^n \quad \text{invariant set for mode } q
\]
\[
E \subseteq Q \times Q \quad \text{set of edges}
\]
\[
G : E \to G \subseteq \mathbb{R}^n \quad \text{guard condition}
\]
\[
R : E \times \mathbb{R}^n \to \mathbb{R}^n \quad \text{reset map}
\]

Considering Equations (1)-(4) for DCCS the set \( \text{Init} = (q_0, x_0) \) is a singleton. Furthermore the guard conditions are defined as
\[
G(q,e) = \{ x : \exists t \text{ such that } \phi(x,q,t,e) = 0 \}
\] (10)
These sets correspond to switching surfaces, which restrict the invariant sets
\[
I(q) = \{ x : \phi(x,q,t,e) < 0, \quad \forall t,e \}.
\] (11)
The trajectory \( x(t) \) will never leave these possibly unbounded sets \( I(q) \) before switching to another mode. Since the state trajectory is continuous at the switching instants, the reset map corresponds to the identity map.

### 2.2. Embedded map and return map of a DCCS

As DCCS produce infinitely long switching sequences with finite activation durations \( \bar{T}(k) = \ell(k+1) - \ell(k) \) and since \( f \) and \( h \) are known, sampling the trajectories at switching instants provides sufficient information to uniquely reconstruct the behaviour between these instants. Hence the switching sequence
\[
\text{Traj}(x_0, q_0, t_0) = \left\{ \begin{array}{l}
\{ \begin{array}{l}
\bar{x}(0) \\
\bar{q}(0) \\
\bar{t}(0)
\end{array}
\end{array}, \begin{array}{l}
\{ \begin{array}{l}
\bar{x}(1) \\
\bar{q}(1) \\
\bar{t}(1)
\end{array}
\end{array}, \ldots
\end{array}
\right.
\]
(12)
is considered as the discrete-event trajectory of the system. The vector \( \bar{s}(k) := (\bar{x}(k), \bar{q}(k), \bar{t}(k)^e) \in S \) is called hybrid switch point and \( S \) is the set of all possible hybrid switch points. From the hybrid model given by Equations (1)-(4) the so called embedded map is derived.

**Definition 1** The embedded map of a DCCS is a function
\[
H : \bar{S} \to \bar{S}
\] (13)
which determines the next hybrid switch point \( \bar{s}(k+1) \) from the current one.

For simple dynamics and event functions, an analytic expression for \( H \) can be obtained. Otherwise the map has to be calculated numerically. By use of the embedded map the analysis of the overall system can be simplified.

By successive application of the embedded map a first return map is obtained, which helps in the analysis of periodic or chaotic behaviour. In the following definition let \( m \in \mathbb{N} \) be a new index.

**Definition 2** The return map of a DCCS is a function
\[
P : \tilde{S} \to \tilde{S}
\]
\[
\bar{s}(m+1) = P(\bar{s}(m)) = H^{(p)}(\bar{s}(m)) = H(\ldots( H(\bar{s}(m))) \ldots),
\] (14)
with \( p \) being the minimal number such that \( \bar{q}(m+1) = \bar{q}(m) \), relates a hybrid switch point \( \bar{s}(m) \) to the next hybrid switch point \( \bar{s}(m+1) \), at which the same mode of operation is first re-activated.

Note that this function is only defined for switch points \( \bar{s} \) for which a cyclic switching sequence is generated by the DCCS. For a bimodal system, the return map is just obtained by \( P(\bar{s}(m)) = H^{(2)}(\bar{s}(m)) = H(H(\bar{s}(m))) \).
Similar to the embedded map, analytic expressions for the return map may be obtained for special cases. For notational convenience both maps (13) and (14) can be decomposed into three components

\[ H = (H_x, H_q, H_r)^T, \quad P = (P_x, P_q, P_r)^T. \] (15)

3. Power converters modelled as DCCS

Fig. 3 depicts a boost converter and its hybrid model corresponding to section 2.1. The power stage is the continuous subsystem with \( x = (i_L, v_o)^T \). In continuous conduction mode the event generator exactly matches with the generator of the implemented current programmed mode control scheme used in the example and the discrete subsystem is exactly the flip flop that controls the power switch. The discrete states are \( q = 1 : S \) closed and \( q = 2 : S \) open. The events are \( e_1 : i_L(t) = i_{ref}, \quad e_2 : \tau = T \) where \( T \) is the period length of the clock generator and \( \tau = t - \max_{k \in \mathbb{Z}} [\bar{T}(k)] \) is the difference between the actual time and the last event occurrence time.

The event function (5) for this control strategy is

\[
\Phi(x(t), q(t), \tau) = \begin{cases} 
\bar{i}_{ref} - i_L(t), & \text{if } q = 1 \\
\tau + \bar{T}(k) - T, & \text{if } q = 2
\end{cases}
\]

The discontinuous conduction mode can be modeled by \( q = 1 : S \) closed, \( D \) blocking, \( q = 2 : S \) open, \( D \) conduct- ing, \( q = 3 : S \) open, \( D \) blocking. Then the events are \( e_1 : i_L(t) = i_{ref}, \quad e_2 : \tau = T, \quad e_3 : i_L(t) = 0. \)

\[
\Phi(x(t), q(t), \tau) = \begin{cases} 
\bar{i}_{ref} - i_L(t), & \text{if } q = 1 \\
\min(\tau + \bar{T}(k) - T, -i_L(t)), & \text{if } q = 2 \\
\tau + \bar{T}(k) + \bar{T}(k-1) - T, & \text{if } q = 3
\end{cases}
\]

Fig. 4 depicts \( P_x \) in mixed continuous discontinuous con-

\[
\begin{array}{ccc}
\bar{q}(k+1) & \bar{q}(k) & e(k+1) \\
2 & 1 & e_1 \\
1 & 2 & e_2 \\
1 & 3 & e_2 \\
3 & 2 & e_3
\end{array}
\]

4. Applications of the Return Map

4.1. Analysis of Transient Behaviour

Transient analysis is used to rate the qualitative behaviour of a system. By iteration of the return map a sampled trajectory of the system state variables arises. [3] shows how to use the return map of a DC-DC converter to investigate the transient behaviour.

4.2. Analysis of the Asymptotic Behaviour

4.2.1. Stability Analysis

The goal of control design in hybrid systems usually is to obtain a stable limit cycle. A stable limit cycle in continuous time domain corresponds to a stable fixed point of the return map of the system or an iterate of the return map. So small signal system stability investigation results in small signal investigation of \( P \) around the fixxed point \( s_{\infty} \). This can be done by use of the Jacobi matrix [4]

\[
J_{\infty} = \frac{\partial P(s)}{\partial s}_{s_{\infty}}.
\] (16)

If all eigenvalues of \( J \) are inside the unit circle the fixxed point and hence the corresponding limit cycle is stable. If one eigenvalue of \( J \) is outside the unit circle the fixxed point and hence the corresponding limit cycle is unstable.

4.2.2. Bifurcations

If the stability of fixed points changes at a specific system parameter value this value is called a bifurcation point. Depending on the case what exactly happens with the fixed points different types of bifurcation are distinguished. Examples are border collision, period doubling and saddle

4.2.3. Chaos

Chaos occurs when all fixed points of the return map and its iterates are unstable i.e. when one eigenvalue of each fixed point is outside the unit circle. Several publications of hybrid systems operating chaotically exist. Analysis usually is based on statistical properties of the signals or on the eigenvalues of the return map.

One side effect of chaotic behaviour is that the power density spectrum (PDS) of the continuous states x becomes continuous. Several groups use this fact for EMC compliance purposes. This will be explained in the next sections.

However in many conventional applications chaotic operation is not desired. So the knowledge about occurrence of chaotic behaviour mostly is used to suppress chaotic operation.

4.2.4. Zeno Behaviour

Zeno behaviour is a phenomenon thoroughly investigated in hybrid systems theory. A system exhibits a Zeno behaviour if it has an infinite number of switchings in a finite time interval. Zeno behaviour is indicated in the return map by a stable fixed point \( \bar{s}_\omega = P(s_\omega) \) with

\[
P_f(s_\omega) = 0.
\]

4.3. Statistical Analysis

Statistical analysis is used for system behaviour analysis in chaotic mode. Important statistical values are the moments of \( x \) [7] and \( f \) e.g. to rate the ripple of continuous variables and the mean switching frequency respectively. Another important statistical value is the power density spectrum of \( x \).

4.3.1. Power Density Spectrum of \( x \)

The power spectrum and power density spectrum (PDS) of the continuous states are important criteria e.g. to rate the EMC compliance of the system [8] or to rate audible components appearing in hybrid systems.

Improvement in audible components suppression or EMI suppression usually is rated by calculating the power within a frequency interval [3].

Approaches to limit this power are passive and active filtering of \( x(t) \), pulse shaping (designing the switching waveform) and modulation of the system by the application of a modulating signal to an additional input of \( \Phi \). The modulation can be classified into periodic [9] and nonperiodic [10] modulation schemes.

In either case the power spectrum or PDS respectively has to be calculated. A convenient method for obtaining the PDS is by using a filtered pulse process [8]. The event times of the pulse process are obtained by \( H_x(s(k)) \) and the pulse intensities and filter function can be obtained from \( f \) for special continuous subsystems.

5. Conclusions

In the paper the class of DCCS was mentioned together with its essential properties. The primary design objectives for such systems were outlined and models of different abstraction levels were proposed. The classification of this class within the existing modeling frameworks for hybrid system was shown. Several analysis problems which represent recent directions of research were mentioned and the use of the embedded map and return map for these analysis problems was shown.

References


Classification and Analytical Method of Switched Dynamical Systems

Yue Ma\textsuperscript{a}, Soumitro Banerjee\textsuperscript{b}, Takuji Kousaka\textsuperscript{c} and Tetsushi Ueta\textsuperscript{d}

\textit{a} Dept. of Electrical and Electronic Engineering, The University of Tokushima, Japan
mayork@gmail.com

\textit{b} Dept. of Electrical Engineering, Indian Institute of Technology, India
soumitro@iitkgp.ac.in

\textit{c} Dept. of Electrical and Electronic Engineering, Oita University, Japan
kousaka@gmail.com

\textit{d} Department of Information Science, Tokushima University, Japan
tetsushi@is.tokushima-u.ac.jp

Abstract—Switched dynamical systems are very common in engineering. Many studies have been done on investigation of their dynamics and bifurcation problems. However, there is still a lack of general classification on this type of systems. Moreover, methods were often developed case by case without general capability. In this article, we propose a new classification method according to properties of sub-systems and switching borders. Furthermore, the approach to apply each kind of switched dynamical systems are summarized. Using this unified method, we can easily find the periodic solution and determine the its stability, then solve complicated bifurcation problems. Special emphasis will be put on those with non-autonomous sub-systems.

1. Introduction

Switched dynamical systems have been subjected to intense investigation in recent years, because such systems find application in a number of engineering areas such as power electronics [1, 2] and mechanical engineering [3]. Switched dynamical systems are defined by two or more sets of differential equations (linear or nonlinear), and the system switches from one set to another when some conditions are satisfied, i.e., the switching borders are researched. For a switched dynamical system, it is the most important and basic problem to find the periodic solution(s) and to determine the stability property of specific solution. Because this problem is the root of all kinds of extended issues such as bifurcation analysis, complex phenomena, control strategy, and so forth.

To solve a periodic solution and determine its stability may seem simple, specially for those traditional continuous dynamical systems for which a nearly complete theory has been established, but is actually quite a complex task for switched dynamical systems. Many methods have been proposed. Most of them, however, are case studies which can only deal with a particular situation. In our earlier studies [4, 5, 6], we proposed a method based on solution flow and its variational equations. By writing equations of solution flows in successive time intervals and then solving them with appropriate numerical methods, we can obtain the periodic solution of switched dynamical systems. Then we define a discrete-time map and investigate the local linear behavior around the fixed point, obtained from the reformation of variational equations. From above process, stability can be studied easily.

But, our earlier studies are also described in some limited conditions, because there was no good classification done for switched dynamical systems. In this article, we will propose a new way to classify switched dynamical systems. This classification approach can cover all kinds of systems and integrate well with our analytical methods. We will extend our previews methods to new classifications. This provides a unify framework for systematic analysis of switched dynamical systems.

2. Classification and system description

2.1. Classification

Switched dynamical systems are consist of two or more continuous dynamical sub-systems and border functions (discrete events) which define the switching condition between those sub-systems. According to if it contains the time explicitly in the sub-systems and the border functions, we can classify switched dynamical systems into four types as following.

1. The sub-systems are autonomous (without the time contained explicitly), and the switching manifold is defined in terms of a simple function of the states, as $f(x) = 0$. Example: the Alpazur oscillator [4].

2. The sub-systems are autonomous, but the switching condition is defined by a function of the states and a periodic function of time. This case was called as “moving border” in [5]. Example: PWM controlled switching power dc/dc converters [1][5], where the switching is controlled by an externally generated triangular wave.

3. The individual sub-systems are non-autonomous, and the border is defined simply as an algebraic equation.
involving the states. For example, mechanical impact oscillators [3].

4. The sub-systems are non-autonomous, and the switching condition has explicit dependence on time. Example: ac/dc power converter, forced neuron [7], bi-oscillators and many complex natural models.

2.2. System description

Type-1 and Type-2 systems are usually described by equations of the form

\[
\begin{align*}
S_1 & : \dot{x} = f_1(x, p_1), & \text{if } x \in M_1, \\
S_2 & : \dot{x} = f_2(x, p_2), & \text{if } x \in M_2,
\end{align*}
\] (1)

where \( x \in \mathbb{R}^n \) are the state variables, \( p_{1,2} \) are the system parameters, \( f_1 \) and \( f_2 \) are the governing functions of sub-system \( S_1 \) and \( S_2 \) when the solution moves in sub-space \( M_1 \) and \( M_2 \) respectively. The functions \( f_1 \) and \( f_2 \) are independent of time.

In contrast, systems of Types 3 and 4 are described by equations of the form

\[
\begin{align*}
S_1 & : \dot{x} = f_1(t, x, p_1), & \text{if } x \in M_1, \\
S_2 & : \dot{x} = f_2(t, x, p_2), & \text{if } x \in M_2,
\end{align*}
\] (2)

In this case the functions \( f_{1,2} \) contain time explicitly.

We assume that there are two sub-spaces \( M_1 \) and \( M_2 \). For each sub-space, “border(s)” exist to make state jump to the other one. For Type-1 and Type-3 systems, the border is defined only by states:

\[
B = \{ x \in \mathbb{R}^n : \beta(x) = 0 \}.
\] (3)

where \( \beta(x) \) is the border function. In systems of Type-2 and Type-4, the border function \( \beta \) has explicit dependence on time, that is,

\[
B = \{ x \in \mathbb{R}^n : \beta(x, t) = 0 \}.
\] (4)

We can summarize above classification method in Fig. 1 clearly.

\[\text{Figure 1: An illustration of classification of switched dynamical systems.}\]

3. Solution flow and fixed point

Obviously, only Type-1 systems are completely autonomous, which is the simplest case. In such systems the discrete-time map has to be obtained by placing a Poincaré surface of section in the state space. Though Type-2 systems have autonomous sub-systems, they are non-autonomous as a whole. Therefore the discrete-time map has to be obtained by stroboscopic sampling in synchronism with the periodicity of the function that defines the switching manifold.

Type-3 systems have non-autonomous sub-systems, and also are non-autonomous as a whole. If the periodicity of the forcing function in all the sub-systems is the same, it is convenient to obtain the discrete map by stroboscopic sampling in synchronism with the periodic input. Type-4 systems can be very complicated if the periodicities of the switching manifold and that of the sub-systems are not identical. Fortunately, however, in most practical systems they are the same, and hence the definition of the stroboscopic sampling is unambiguous. For those cases where two frequencies are different, two types of map are often considered.

The generalized methods of following periodic orbits and obtaining the Jacobian matrix of Type-1 and Type-2 systems have been developed earlier [4][5]. As the logical next step, in this paper we develop the method for handling systems of Type-3.

For Type-1 and Type-3, the two sub-spaces are given by

\[
M_1 = \{ x \in \mathbb{R}^n : \beta(x) > 0 \}.
\]

\[
M_2 = \{ x \in \mathbb{R}^n : \beta(x) < 0 \}.
\] (5)

The solution of the system is governed by the functions \( f_1 \) and \( f_2 \) as given in (1) or (2), depending on whether the state is in \( M_1 \) or in \( M_2 \). We express the solution functions as

\[
\begin{align*}
x(t) &= \varphi_1(t, t_0, x_{01}), & \text{if } x \in M_1, \\
x(t) &= \varphi_2(t, t_0, x_{02}), & \text{if } x \in M_2,
\end{align*}
\] (6)

where \( x_{01} \) and \( x_{02} \) are the initial conditions, and \( t_0 \) is the starting time of the flow.

For Type-3 systems, we assume that both \( f_1(t, x, p_1) \) and \( f_2(t, x, p_2) \) are periodic functions with identical period \( T \), i.e.,

\[
f_{1,2}(t, x, p_{1,2}) = f_{1,2}(t + T, x, p_{1,2}).
\] (7)

Now, define the stroboscopic mapping as

\[
F_m : \mathbb{R}^n \rightarrow \mathbb{R}^n; \quad x(0) \mapsto x(T).
\] (8)

In this map, if a solution satisfies \( x(0) = x(kT), k = 1, 2, \cdots \), it is a period-\( k \) solution.

For Type-3 systems, the sampling points will not generally lie on the switching manifold. Suppose there is a solution flow in one period as follows.

1. Starting at \( x_0 \in M_1 \), the state moves in \( M_1 \) until the solution satisfies the switching condition \( B \) at \( x_1 \) after time \( \tau_1 \);
2. Crossing the border $B$, the state moves into $M_2$. Then at time $\tau_2$, the solution flow crosses the switching manifold again at $x_2$, and returns to sub-space $M_1$.

3. Finally, the state reaches $x_1$ after time $T$.

According to (6) and (3), the above solution flow can be described as

$$
\begin{align*}
 x_1 &= \varphi_1(t_1, 0, x_0) \quad (9) \\
 x_2 &= \varphi_2(t_2, \tau_1, x_1) \quad (10) \\
 x_3 &= \varphi_1(T - \tau_2, 0, x_2) \quad (11) \\
 \beta(x_1) &= 0 \quad (12) \\
 \beta(x_2) &= 0 \quad (13)
\end{align*}
$$

This is shown schematically in Fig. 2.

In the earlier work on Type-1 and Type-2 systems, the initial time in (10) and (11) were taken as zero. That means they can be reduced to the form

$$
\begin{align*}
 x_2 &= \varphi_2(t_2 - \tau_1, 0, x_1) \quad (14)
\end{align*}
$$

where only time differences were considered. However in Type-3 systems, the value of initial time does matter, because the functions $f_1$ and $f_2$ have explicit dependence on time. Therefore the above reduced forms are not applicable. This feature makes the handling of Type-3 systems different from that of Type-1 and Type-2 systems.

The behavior of periodic solution flow of the system is reduced to the behavior of fixed points of the map, $F_m : x_0 \mapsto x_1$. In this section, we will introduce the method to obtain the location of the fixed point and to analyze its stability property.

For systems of Type-3, a fixed point exists if $x(0) = x(T)$, i.e., if $x_0 = x_1$. There are $3n + 2$ scalar equations in (9)-(13), which can be solved using appropriate numerical technique to obtain the $3n + 2$ scalar unknowns $\{x_0, x_1, x_2, \tau_1, \tau_2\}$. We have used the Newton’s method for this purpose. To apply the Newton’s method, we need to know the partial derivatives of the solution with respect to the unknowns, for which the Jacobian matrix of the solution flow is necessary.

Because we are considering the general case, we cannot assume that the solutions can be obtained in closed form. We therefore have to obtain the partial derivatives using numerical method.

4. Stability of the fixed point

From (9), (10) and (11) we can write

$$
\begin{align*}
 \frac{\partial x_1}{\partial x_0} &= \frac{\partial \varphi_1}{\partial t} \bigg|_{t=\tau_1} \frac{\partial \varphi_1}{\partial x_0} + \frac{\partial \varphi_1}{\partial \tau_1} \frac{\partial \varphi_1}{\partial x_0}, \\
 \frac{\partial x_2}{\partial x_0} &= \frac{\partial \varphi_2}{\partial t} \bigg|_{t=\tau_1} \frac{\partial \varphi_2}{\partial x_0} + \frac{\partial \varphi_2}{\partial \tau_1} \frac{\partial \varphi_1}{\partial x_0} + \frac{\partial \varphi_2}{\partial \tau_2} \frac{\partial \varphi_1}{\partial x_0} + \frac{\partial \varphi_2}{\partial \tau_2} \frac{\partial \varphi_2}{\partial x_0}, \\
 \frac{\partial x_3}{\partial x_0} &= \frac{\partial \varphi_1}{\partial \tau_1} \frac{\partial \varphi_1}{\partial x_0} + \frac{\partial \varphi_1}{\partial \tau_2} \frac{\partial \varphi_2}{\partial x_0} + \frac{\partial \varphi_1}{\partial \tau_2} \frac{\partial \varphi_2}{\partial x_0} + \frac{\partial \varphi_2}{\partial \tau_2} \frac{\partial \varphi_2}{\partial x_0}
\end{align*}
$$

where we have used the abbreviations like $\partial \varphi_1 / \partial x_0$ to denote $\partial \varphi_1 / \partial x_{k-x_0}$, etc.

Now, partial derivatives of the solution flow with respect to time can be written as

$$
\begin{align*}
 \frac{\partial \varphi_1}{\partial t} \bigg|_{t=\tau_1} &= f_1(\tau_1, x_1, p_1), \\
 \frac{\partial \varphi_2}{\partial t} \bigg|_{t=\tau_2} &= f_2(\tau_2, x_2, p_2). \quad (20)
\end{align*}
$$

On the other hand, partial derivatives of the solution flow with respect to the initial time points $\tau_1$ and $\tau_2$ [8] are

$$
\begin{align*}
 \frac{\partial \varphi_2}{\partial \tau_1} &= -\frac{\partial \varphi_2}{\partial x_1} f_2(\tau_1, x_1, p_2), \\
 \frac{\partial \varphi_1}{\partial \tau_2} &= -\frac{\partial \varphi_1}{\partial x_1} f_1(\tau_2, x_2, p_1).
\end{align*}
$$

This is the key point by which Type-1&2 different from Type-3&4. Because in Type-1&2, initial times of each flow are always regard as zero, and no derivative is needed. Substituting these into (15), (16), and (17) we get

$$
\begin{align*}
 \frac{\partial x_1}{\partial x_0} &= f_1(\tau_1, x_1, p_1) \frac{\partial \tau_1}{\partial x_0} + \frac{\partial \varphi_1}{\partial x_0}, \\
 \frac{\partial x_2}{\partial x_0} &= f_2(\tau_2, x_2, p_2) \frac{\partial \tau_2}{\partial x_0} + \frac{\partial \varphi_2}{\partial x_0} \\
 \frac{\partial x_3}{\partial x_0} &= -\frac{\partial \varphi_1}{\partial \tau_1} f_2(\tau_1, x_1, p_2) \frac{\partial \tau_1}{\partial x_0} - \frac{\partial \varphi_2}{\partial \tau_2} f_1(\tau_2, x_2, p_1) \frac{\partial \tau_2}{\partial x_0} + \frac{\partial \varphi_1}{\partial x_0}
\end{align*}
$$

Note that (24) gives the Jacobian matrix of the Poincaré map $x_0 \mapsto x_3$, which will be used to analyze the stability property.

In order to evaluate (22), (23), and (24), we need to obtain the values of $\partial \varphi_1 / \partial x_0$, $\partial \varphi_2 / \partial x_0$, and $\partial \varphi_1 / \partial \tau_1$. These partial derivatives of the solution functions with respect to the initial conditions are obtained by solving the differential equations [8]

$$
\frac{d}{dt} \left( \frac{\partial \varphi_1}{\partial x_0} \right) = \frac{\partial f_1}{\partial x} \left( \frac{\partial \varphi_1}{\partial x_0} \right), \quad \frac{\partial \varphi_1}{\partial \tau_1} \bigg|_{t=0} = I^n
$$

65
Finally, to evaluate (22), (23), and (24), we also need the partial derivatives of the evolution within each sub-system with respect to the initial condition. Since the times of evolution within each sub-system is determined by the switching condition, these partial derivatives can be obtained from the expressions of the switching manifolds. From (12) and (13) we can write

$$\frac{\partial \beta(x_1)}{\partial x_{0}} = \frac{\partial \beta(x_2)}{\partial x_{0}} = 0,$$

and

$$\frac{\partial \beta(x_2)}{\partial x_{0}} = \frac{\partial \beta(x_1)}{\partial x_{0}} = 0.$$  

(28)

Because border function $\beta(x)$ is 0 is a 1 dimensional equation, $\partial \beta / \partial x$ is a $1 \times n$ matrix. So, when $\beta(x)$ is in a simple form like $x(1) = \text{constant}, \partial \beta / \partial x = [1, 0, \cdots, 0]$. Thus, if $\partial \beta / \partial x \cdot \partial x_{0} = 0$ is given by (28), the corresponding first row of $\partial x_{1}/\partial x_{0}$ must be $[0, \cdots, 0]$. But if the switching manifold is given by an equation involved more than one state variable like the form $\beta(x) = x(1) - x(2) = 0$, then $\partial \beta / \partial x = [1, -1, 0, \cdots, 0]$. Thus the difference between first two rows of $\partial x_{1}/\partial x_{0}$ will be zero. Anyway, $\partial \beta / \partial x$ supplies the additional information of specific dimension(s) related to border function, by which we can solve the partial derivatives of the times of evolution with respect to the initial state $x_0$. Substituting (22) and (23) into (28), we get

$$\frac{\partial r_1}{\partial x_0} = -\frac{\partial \beta}{\partial x} \frac{\partial \phi_1}{\partial x_{0}}, \quad \text{and}$$

$$\frac{\partial r_2}{\partial x_0} = -\frac{\partial \beta}{\partial x} \frac{\partial \phi_2}{\partial x_{0}} = \frac{\partial \beta}{\partial x} \frac{\partial \phi_2}{\partial x_{1}} \frac{\partial x_{1}}{\partial x_{0}} = \frac{\partial \beta}{\partial x} \frac{f_2(t_2, x_2, p_2)}{f_1(t_1, x_1, p_1)} \frac{\partial r_1}{\partial x_{0}}.$$  

(29)

These expressions, substituted into (24), give the Jacobian matrix $\partial x_{1}/\partial x_{0}$ of the Poincaré map. Then by finding the roots of the characteristic equation, we can determine the stability of specific fixed point, i.e., corresponding periodic solution.

Certainly, the system may exhibit more complicated orbits with higher periodicities and larger number of transitions between $M_1$ and $M_2$. In those cases, the solution flow can be formulated according to the time instants at which the flow crosses the border or reaches the sampling time $nT$. The process will only involve a larger number of equations of the form (9)-(13). Therefore, to compute a complicated orbit we will only need to solve a larger number of simultaneous algebraic equations for a larger number of unknowns. An effective numerical method can solve them and can find the position of the fixed point without much difficulty. Note that for fixed points of period-2 or higher, the Jacobian matrix is to be composed by multiplying the Jacobian matrices of the individual submappings that take the state from the start of one sub-system to the start of the next sub-system, finally to the end of the periodicity. Moreover, all the terms used to get the Jacobian matrix are derived in the process of finding the fixed point. Thus, no additional calculation is needed to determine the stability.

5. Conclusion

In this paper, we proposed a method to classify and analyze the switched dynamical systems. One of the types was introduced in detail. The approach is general and easy to apply.

References


Mixed-Signal Modeling and Design Aspects of Clock-Recovery Circuits and Continuous-Time Sigma-Delta Modulators

F. A. Probst, J. Schmackers and W. Mathis†

†Institute of Theoretical Electrical Engineering, Department of Electrical Engineering and Computer Science
University of Hannover, Germany
Email: probst@tet.uni-hannover.de, schmackers@tet.uni-hannover.de, mathis@tet.uni-hannover.de

Abstract—In this paper we consider the application of nonlinear mixed signal modeling equations for advanced integrated analogue circuits and study solution methods for these equations. Based on these results we analyse current design concepts and discuss improvements with respect to insert nonlinear modeling aspects.

1. Introduction

During the last decade RF microelectronics has been focusing on so-called mixed-signal systems where the functionality of the system can be divided into digital (binary) and analog operating parts. Typical examples are clock and data recovery circuits (CDRs) and continuous-time (CT) Sigma-Delta (Σ∆) modulators. CDRs are widely used in receiver architectures of optical communications performing the task of recovering the clock from noisy serial data and using this clock to sample the data providing noise free binary signals for further processing. Applying high oversampling ratios and noise shaping principles Σ∆ modulation can achieve high resolution A/D and D/A conversion for low bandwidth demands although using a comparator for 1 bit quantization [1],[2]. Further application areas are digital audio signal processing and digital telephony. As mentioned above CT Σ∆ modulators and CDRs include analog and digital elements so that an accurate description simply with differential equations is difficult. However a pure digital description is also not useful if an accurate model is needed. Therefore mixed-signal models should include CT as well as discrete-time (DT) equations which means we have difference-differential as well as difference-integral equations.

Studying the modeling aspects of CDRs and CT Σ∆ modulators we will show that a unified approach to mixed-signal systems seems to be impossible. However we present some ideas to describe such types of systems. For this purpose we analyse conventional modeling approaches of CDRs and CT Σ∆ modulators and present some advanced ideas for mixed-signal modeling. Finally we discuss these approaches in relation to system design.

2. Mixed Signal Modeling: Conventional Methods

2.1. PLL-Based CDR

In this Section we consider CDR architectures. Obviously CDRs are included in the class of mixed-signal systems. Mainly there are two methods of realising the phase detector (PD) in CDRs [3]. Since the incoming data \( D_{in} \) does not have a frequency component at the clock frequency \( f_{CLK} \), simple PDs such as multipliers or XOR gates that are used for PLLs do not suffice. For this reason, the PD needs to perform edge detection in order to correctly detect the frequency. There are two main types of PDs that meet this criterium, the linear PD (e.g. Hogge PD) and the so called bang-bang PD (e.g. Alexander PD). In a linear PD, the resulting VCO control voltage \( V_{VCO} \) consists of pulses linearly dependent on the phase difference. An ideal bang-bang characteristic only determines the sign of the phase difference.

![Figure 1: CDR Structure](image-url)

2.1.1. Overview of Modeling Approaches

Both linear and bang-bang PDs are usually implemented using flip-flops and therefore introducing a digital element into the circuit. From a CT point-of-view, the phase detector is therefore highly nonlinear. This means that the description of the phase detector is the challenging part in modeling the CDR circuit. The mixed-signal model presented in Section 3.1 is a model for an Alexander PD that is an adaptation of the model presented in Section 2.1.2 which characterizes a linear PD. Whereas the model for the linear PD uses the phase error between \( D_{in} \) and \( V_o \) (see Figure 1), the model for the Alexander PD uses the structure of the PD to derive conditions for the charge current.
i_C. i_C can either be positive, negative or off. The VCO is modeled as an integrator and the filter by its impulse response in both approaches.

One possibility to describe the bang-bang PD more accurately is to include effects such as metastability of the flip-flops and jitter on the incoming data. Modeling input jitter and metastability introduces a linear relation between the phase error $\Delta \phi$ and the VCO control voltage $V_{VCO}$ for small phase errors [4].

2.1.2. Model for the Linear Phase Detector

A method for modeling linear PDs such as the Hogge PD consists of including the sampling nature of the charge pump in the model [5],[6]. The input phase difference $\theta_i$ can be described as follows if the input signal is sinusoidal:

$$\theta_i(t) = \theta_i(0) + \omega_i t,$$

$$\theta_o(t) = \theta_o(0) + \omega_o t + K_{VCO} \int_0^t V_{VCO}(\tau)d\tau,$$

$$\theta_e(t) = \theta_i - \theta_o.$$ (3)

Here $\omega_i$ is the input frequency, $\omega_o$ is the free running frequency of the VCO and $K_{VCO}$ is the VCO gain. For positive phase difference $\theta_e$, the VCO control voltage $V_{VCO}$ needs to be increased and decreased for $\theta_e < 0$. Therefore the charge current $i_C$ can be described with

$$i_C = \begin{cases} 
I_{CP} \cdot \text{sgn}(\theta_e) & \text{if } 0 \leq t \leq t_p(\theta_e) \\
0 & \text{if } t_p(\theta_e) < t < T_-(\theta_e), f_{CLK}, f_{in} \\
< T_-(\theta_e), f_{CLK}, f_{in} 
\end{cases}$$ (4)

where $t_p$ and $T_-$ are illustrated in Figure 2 and $I_{CP}$ represents the constant current provided by the charge pump. In Figure 2 $D_{in}$ represents the input data whereas $D_{in,f}$ is its clock frequency and $V_{CLK}$ represents the output of the VCO. The filter is described by its impulse response $f$:

$$V_{VCO} = f * i_C.$$ (5)

Substituting Eq. (5) and Eq. (4) in Eq. (2), the loop is closed and a complete description of the system is available.

2.2. Continuous-Time Sigma-Delta Modulator

An overall model for all aspects of $\Sigma\Delta$ modulation cannot be found because of the nonlinearity inside a feedback system. There are mainly two conventional modeling approaches used for $\Sigma\Delta$ modulation. The most established modeling method replaces the nonlinear quantizer by an additive signal which is formed by the difference between the output and the input of the quantizer. This quantization error is often called quantization noise and is assumed to be white. Another main issue of $\Sigma\Delta$ modulation is system stability.

2.2.1. Additive, white noise approximation

Replacing the nonlinear deterministic by a linear stochastic system it is possible to give a qualitative explanation of the system’s behaviour. Further it is possible to approximately quantify the resolution of the conversion by calculating a signal-to-noise-ratio (SNR). Two independent transfer functions can be calculated describing the influence of the quantization noise and the signal input to the output bitstream propagating the possibility to shape the noise totally independent from the input signal.

The question arises whether the quantization noise is actually white. Gray [7] was the first who demonstrated mathematically that especially for $\Sigma\Delta$ modulation the quantization noise is not white at all. By developing a Fourier series of the quantization noise he deduced several characteristic statistical properties for DC-an sinusoidal inputs.

Furthermore the white noise approximation is not able to answer stability issues due to the lack of internal dynamic behaviour description. Conventional stability discussions are motivated by linearizing methods such as root locus or describing function methods [1],[2] but also cannot include nonlinear effects.

2.2.2. Transformation between continuous-time and discrete-time Sigma-Delta modulators

Established designing methods of CT $\Sigma\Delta$ modulators use DT designing approaches [1]. After designing a DT loop filter the resulting transfer function is transformed into an equivalent CT transfer function. CT and DT $\Sigma\Delta$ modulators are meant to be equivalent if they produce the same output bitstreams for the same input signals. To assume the same output bitstream...
the comparator input has to have the same sign at sampling times. The general approach is to cut of the loop at the quantizer and to postulate that the impulse responses of the DT and the CT modulator from the quantizer output to its input have to be the same at sampling times [1]. This idea is illustrated in Figure 3. According to impulse invariance the following condition has to be satisfied

$$Z^{-1}\{H(z)\} = \mathcal{Z}^{-1}\left\{F_{DA}(s)\hat{H}(s)\right\}_{t=n\Delta t},$$

(6)

where \(H(z)\) denotes the DT and \(\hat{H}(s)\) the associated CT transfer function, \(\mathcal{L}\{\cdot\}\) the Laplace-Transformation, \(\mathcal{Z}\{\cdot\}\) the Z-Transformation and \(F_{DA}(s)\) the transfer function of the DAC in the feedback loop. Disregarding the influence of the filtered modulator input signal on the quantizer input the condition of equivalent output bitstreams cannot be satisfied. Despite the fact that the approximation of an CT integration by a DT summation is improved by the small sampling time the discontinuity of the nonlinearity cannot allow even small variations of the quantizer inputs. Due to the closed loop structure of the \(\Sigma\Delta\) modulator the system tries to hold the quantizer input at zero which at the same time is the switching threshold of the quantizer. Different loop filters result in different types of disturbances caused by the modulator input signal. Hence it is impossible to find equivalent CT and DT \(\Sigma\Delta\) modulators.

3. Generalized Mixed Signal Modeling Approaches

3.1. PLL-Based CDR

The method described in Section 2.1.2 can be transferred to Alexander PDs (bang-bang type) as well. Eq. (4) needs to be adapted so that the conditions for the charge pump current are

$$i_C = \begin{cases} I_{CP} & \text{if } y = 1 \\ -I_{CP} & \text{if } x = 1 \\ 0 & \text{if } x = 0 \text{ and } y = 0, \end{cases}$$

(7)

where \(x\) and \(y\) are the UP and DOWN outputs of the Alexander PD. In this description the phase difference \(\theta_0\) does not occur which is characteristic for bang-bang PDs. A relation between the data input \(D_{in}\), the sampling time of the flip-flops in the PD and the PD outputs \(x\) and \(y\) is needed. To achieve this, the D-flip-flop input-output relation is described as

$$d_{out}(T) = F(d_{in}) = d_{in}(T - 1)$$

(8)

where \(T\) represents the sampling time and \(d_{in}\) and \(d_{out}\) the in- and output data of the flip-flop [8]. The UP and DOWN outputs of the Alexander PD can therefore be modeled with \(d_{in} = D_{in}\). One of the flip-flops in the Alexander PD samples on negative edges. To avoid sampling times of \(T = \frac{1}{2}\) we assume that data is sampled on both positive and negative edges:

$$y(T) = D_{in}(T - 2) \oplus D_{in}(T - 3),$$

(9)

$$x(T) = D_{in}(T - 4) \oplus D_{in}(T - 3).$$

(10)

In a CDR the sampling times \(T = \{T_0, T_1, T_2, \ldots, T_n\}\) are not equidistant as in conventional sequential logic since the clock period is the variable to be controlled. Therefore,

$$T_n = \sum_{i=0}^{n} \Delta T_i = \sum_{i=0}^{n} \frac{t_{CLK,i}}{2}$$

(11)

since the difference \(\Delta T_i\) between sampling times \(T_i\) is half a clock period \(t_{CLK}\).

$$\Delta T_i = \frac{t_{CLK,i}}{2} \neq \text{const.}$$

(12)

With the VCO relation

$$\omega_C = \omega_0 + K_{VCO}V_{VCO}(t)$$

and with \(t_{CLK} = \frac{2\pi}{\omega_C(t)}\) follows

$$t_{CLK} = \frac{2\pi}{\omega_0 + K_{VCO}V_{VCO}}$$

(14)

Eq. (5) and Eq. (7) can now be substituted and the loop is therefore closed.

3.2. Continuous-Time Sigma-Delta Modulator

Usual modeling techniques for \(\Sigma\Delta\) modulators use linearized models in the frequency domain, as seen in Section 2.2. A different approach was introduced in [9], [10] by providing a time domain description and solving the characteristic nonlinear difference-integral equation. Finding a solution is complicated by the nonlinear comparator which can ideally be described mathematically by a Heaviside function. If certain conditions will be satisfied the argument of the Heaviside function can be replaced by the greatest whole number less or equal than the argument of the function.

$$\frac{1}{T_{CLK}} \int_{0}^{T} \frac{1}{T_{CLK}} \int_{0}^{T} \ldots \int_{0}^{T} \left[ \frac{1}{T_{CLK}} \int_{0}^{T} \ldots \int_{0}^{T} \right]_{T_{CLK}}^{T}$$

Figure 4: First Order CT \(\Sigma\Delta\) modulator

The model illustrated in Figure 4 shows a first order \(\Sigma\Delta\) modulator. According to the formula

$$x(t) = 2\sigma\{w(t)\} - 1$$

(15)
the ideal comparator toggles its output \(x(t)\) due to the sign of the input signal, where \(\sigma\) denotes the Heaviside function. The filter integrates the difference between the analog input signal \(u(t)\) and the feedback output bitstream \(x[k]\), weighted by \(\alpha U_{ref}\), resulting in the integrator output

\[
w(t) = \frac{1}{T_I} \int_{t_0}^{t} u(\tau) - \alpha U_{ref} x(\tau) d\tau + w(t_0).
\] (16)

Combining the above equations and setting the initial condition \(w(t_0) = 0\) yields to the following relation for the output bitstream:

\[
x(t) = 2\sigma \left\{ \frac{1}{T_I} \int_{t_0}^{t} u(\tau) d\tau - \frac{\alpha U_{ref}}{T_I} \int_{t_0}^{t} x(\tau) d\tau \right\} - 1. \quad (17)
\]

Due to the sampling inside the comparator the second integral can be replaced by a sum since the output bitstream is a discrete function \(x[k] := x(k\Delta t)\), which is constant during a sampling period \(\Delta t\). Thus Eq. (17) can be written as a nonlinear difference-integral equation

\[
x[k] = 2\sigma \left\{ \frac{1}{T_I} \int_{k\Delta t}^{(k+1)\Delta t} u(\tau) d\tau - \frac{\alpha U_{ref}\Delta t}{T_I} \sum_{j=0}^{k-1} x[j] \right\} - 1. \quad (18)
\]

Applying the definitions for the first difference

\[
y[k+1] - y[k] := x[k] \quad (19)
\]

\[
y[0] = 0 \quad (20)
\]

the difference-integral equation (18) can be solved explicitly. The solution is given by

\[
y[k] = 2 \left( \frac{1}{2\alpha U_{ref}\Delta t} \int_{0}^{(k-1)\Delta t} u(\tau) d\tau + \frac{k - 1}{2} + 1 \right) - k, \quad (21)
\]

with \(\forall k \in \mathbb{N}\) if the condition \(|u(t)| \leq \alpha U_{ref}\) is satisfied and where \(\lfloor \xi \rfloor\) denotes the next whole number less or equal than \(\xi\) [9],[10].

4. Design Aspects and Conclusion

Conventional design strategies mainly use linear models that are unable to describe every phenomenon occurring in the circuit. More generalized models such as the ones presented here allow the designer to better determine the behaviour of the circuit before realising the circuit on transistor level.

CDRs are usually described with linearised baseband models which are also used for PLL modeling. The model described in Section 3.1 provides the possibility of revising the design process so that more phenomena occurring in CDRs are taken into account in an early design step.

Popular design methods of CT \(\Sigma\Delta\) use linearizing models to find a DT filter transfer function which is subsequently transformed into a CT filter as mentioned in Section 2.2.2. In opposition to those linearized frequency domain approaches the time domain model offers the possibility of analysing modulator stability and deducing decimation filter properties [11].

References


Switching Dynamics, Chaos, and Singular Perturbation Analysis of Heart Arrhythmia

Naoyuki Takahashi, Shinji Doi and Sadatoshi Kumagai
Graduate School of Engineering, Osaka University, Suita 565-0871 JAPAN
Email: [nookyui@ic.eei, doi@eii, kumagai@eii].eng.osaka-u.ac.jp

Abstract—The Luo-Rudy model is the one of typical models of ventricular myocardial cells and can be treated as a ‘2-slow’ system rather than a 1-slow system; orbits or solutions can be analyzed in terms of switching dynamics on a two-dimensional slow manifold in a high-dimensional state space. We also show that geometric conformation changes of the slow (switching) manifold lead to anomalous (chaotic) behavior and discuss the relation of this chaotic behavior to heart arrhythmia.

1. Introduction

Action potential plays a crucial role in life, such as brain, nervous systems, heart, pancreas, etc. The action potential is described by the so-called Hodgkin-Huxley (HH) type equations. Biological processes described by such HH-type equations usually have multi-timescales: singularly perturbed dynamical systems. Decomposition of singularly perturbed systems into fast and slow subsystems is very useful to reveal the underlying dynamics of high-dimensional systems. Many biological examples such as bursting cells are usually explained in a framework of 1-slow dynamics [2, 3]. In this paper, we show that the Luo-Rudy model [4] of ventricular myocardial cells can not be treated as a ‘1-slow’ system but can be explained in terms of ‘2-slow’ dynamics. In 2-slow dynamical systems, solution orbits follow two-dimensional slow manifolds in a high-dimensional state space and switch between several branches of a two-dimensional slow manifold. The dynamics of 2-slow dynamical systems can produce essentially different behavior from that of 1-slow dynamical systems since the switching or jumping at a two-dimensional manifold can produce as rich dynamics as that of one-dimensional discrete-time dynamical systems (1D mappings) [1]. Various phenomena including an abnormal behavior which induces a biologically dangerous arrhythmia of heart are clearly explained in the framework of switching dynamics and from the geometrical viewpoint of slow manifold conformation changes. Thus, this paper presents not only an important view on biological/medical sciences but also a significant mathematical example with 2-slow dynamics.

2. Luo-Rudy Model of Ventricular Myocardial Cells

The Luo-Rudy (LR) model [4] of ventricular myocardial cells is described by the following differential equations of eight variables:

\[
\frac{dV}{dt} = -\frac{1}{C_m}(I_{\text{Na}} + I_d + I_K + I_{K1} + I_{K_C} + I_L - I_{\text{ext}}), \quad (1a)
\]

\[
\frac{dy}{dt} = \frac{1}{\tau_y(V)}(y_{\alpha}(V) - y), \quad (y = m, h, j, d, f, X), \quad (1b)
\]

\[
\frac{d[Ca]_i}{dt} = -10^{-4}I_{d_0} + 0.07(10^{-4} - [Ca]_i), \quad (1c)
\]

where the variable \(V\) (mV) is the membrane potential and \(y\) denotes one of the six gating variables \(m, h, j, d, f, X\) (dimensionless). The variable [Ca] denotes the intra-cellular concentration of Ca\(^{2+}\) ions. Thus the LR model has an eight-dimensional dynamics. \(C_m = 1 \mu F/cm^2\) is the capacitance of cell membrane. \(I_{\text{Na}}, ..., I_L\) are the ionic currents which depend on the membrane potential \(V\), the gating variables and the Ca\(^{2+}\) concentration. See Appendix for the exact formulae of these ionic currents and of the functions \(\alpha_y(V)\) and \(\beta_y(V)\).

Figure 1 shows a typical solution of eq. (1): Right panel (b) is the waveform of \(V\) and waveforms of other six gating variables and Ca\(^{2+}\) are shown in (a). For the later references, the ‘depolarized’ (high voltage) and ‘repolarized’ (low voltage) portions of the membrane potential \(V\) are denoted by solid and broken lines, respectively. We note that...
the LR model (1) with normal parameter values has a globally stable equilibrium point when $I_{\text{ext}} = 0$. Thus the depolarization of Fig.1 is induced by the application of an external input pulse: $I_{\text{ext}} > 0$.

3. Slow-Fast Decomposition Analysis of the Luo-Rudy Model of Heart Cells

3.1. Singularly-Perturbed Dynamical Systems

In general, a system with multiple time scales may be denoted as follows:

\[
\frac{dx}{dt} = f(x, y), \quad x \in R^n, \quad y \in R^m, \quad (2a)
\]

\[
\frac{dy}{dt} = \epsilon g(x, y), \quad \epsilon \ll 1. \quad (2b)
\]

Equation (2b) is called a slow subsystem since the value of $y$ changes slowly while eq.(2a) a fast subsystem. The whole equations (2) are called a full system especially. So-called slow-fast analysis divides the full system to the slow and fast subsystems. In the fast subsystem (2a), the slow variable $y$ is considered as a constant or a parameter. The variable $x$ changes more quickly than $y$ and thus $x$ is considered to stay close to the attractor (stable equilibrium points, limit cycles, etc.) of the ‘fast’ subsystem for a fixed value of $y$. Especially, if the attractor of the fast subsystem is an equilibrium point $x^*$, $x(t)$ stays near a slow manifold: $f(x^*, y) = 0$. (Slow manifold is a surface in which the equilibria of the fast subsystem are denoted as a function of slow variable $y$.) Thus, the slow manifold is considered as a ‘bifurcation diagram’ of the fast subsystem as for the bifurcation parameter $y$.

The variable $y$ changes slowly with a velocity $\epsilon g(x, y)$ in which $x$ is considered to be in the neighborhood of the attractor. The attractor of the fast subsystem may change if $y$ is varied. The problem of analysis of the dependence of the attractor on the parameter $y$ is a bifurcation problem. Thus the slow-fast analysis reduces the analysis of the ‘full’ system to the bifurcation problem of the ‘fast’ subsystem with a slowly-varying bifurcation parameter.

It is worth noting that the determination of which variable(s) is slow or fast one in such differential equations as the LR model (1) is not apparent at a glance of the equations; it is determined by the ‘relative speed’ among variables. Thus, we first examine which variable(s) is a slow variable in the next subsection.

3.2. 1-Slow versus 2-Slow Dynamics

Figure 2 illustrates a slow-fast decomposition analysis under an assumption that the LR model could be considered as a 1-slow dynamical system. The thin curves in Fig.2 denote equilibrium points of a (seven-dimensional) ‘fast’ subsystem as a function of a slow variable (i.e. a bifurcation diagram of the fast subsystem with a slow variable considered as a bifurcation parameter). Each panel of the figure considers one of all seven variables other than $V$ in eq. (1) as a slow variable.

For example, the folded thin curve with a label HB in panel (e) is the bifurcation diagram of the fast subsystem where the gating variable $f$ is the slow variable. The equilibrium points of the fast subsystem suffer a Hopf bifurcation at the point labeled by HB; the solid and broken curve denotes stable and unstable equilibria. Thick curve (dots) is the projection of solution orbit of the ‘full’ system to the $f$-$V$ phase plane where black and gray portions correspond to the depolarized and repolarized membrane potential, respectively, shown in Fig.1b (the border is marked by an arrow). In none of seven panels of Fig.2, the orbit of full system follows the equilibria of the fast subsystem (one-dimensional slow manifold), which means that the LR model can not be described as a 1-slow dynamical system. Thus we seek the other possibility that the LR model has two slow variables in the next subsection.

3.3. Switching Dynamics on Two-Dimensional Slow Manifolds

We have examined the LR model as a 2-slow system choosing various combinations of two variables and finally...
found two gating variables \( f \) and \( X \) as a candidate of slow variables. The folded surface (slow manifold) of Fig.3(a) shows the equilibria \( V^*(f,X) \) of the fast subsystem as a function of two slow variables: \( f \) and \( X \) (the surface is connected, although two separate surfaces are seen in this scale.). The see or the folded portion (with SN labeled) corresponds to the saddle-node bifurcation points of the (six-dimensional) fast sub-system. Thus, both upper and lower branches of the folded two-dimensional slow manifold correspond to the stable equilibria of the ‘fast’ subsystem and the middle one the unstable equilibria. The slow manifold \( V = V^*(f,X) \) can be obtained by the following equa-

The orbit (trajectory) of the full system (the LR model (1)) shown in dark curve is also superimposed on Fig.3(a). Solid and dotted portions correspond to the depolarized and repolarized portions of membrane potential shown in Fig.1(b), respectively. The starting point of depolarized portion located in lower-right of Fig.3 marked with a red circle) corresponds to a stable equilibrium point of ‘full’ system. An application of a brief current pulse \( I_{\text{ext}} \) moves the state point of LR model almost vertically apart from the stable equilibrium. After a slight overshoot of this vertical movement, the state point moves down and follows the upper branch of the slow manifold to the lower-left direction. When the orbit hits the see (fold) of the slow manifold, it changes the destination to lower-right direction and leaves the upper branch toward the lower branch, and finally returns to the starting point (a stable equilibrium) following the lower branch of the slow manifold. Thus, various properties of a heart cell’s membrane potential can be analyzed based on the geometrical characteristics of the slow manifold shown in Fig.3.

Figures 3(b) and (c) show the vector field (denoted by arrows) of the slow sub-system:

\[
\frac{df}{dt} = \frac{1}{\tau_f(V^*(f,X))}(f_0(V^*(f,X)) - f),
\]

on the lower and upper branches of the slow manifold, respectively, where \( V^*(f,X) \) is the slow manifold defined in eq. (3) and other functions are described in Appendix. The orbit of full system is also superimposed on this \( f-X \) phase plane. The meaning of solid and dotted portions of this orbit is the same as Figs.1 and 3(a). Thus, we can see that the repolarized portion (dotted curve) of the full system’s orbit moves along the vector field on the lower slow manifold (Fig.3(b)) while the depolarized portion (solid curve) follows the vector field on the upper slow manifold (Fig.3(c)). The red circle shows the (globally) stable equilibrium of the ‘full’ system.

The action potential duration (APD) is a very important factor in heart arrhythmia and corresponds to the solid curve shown in Figs.1 and 3. Thus a quantitative property such as the APD can be also obtained by the flow of the slow subsystem (4) as follows:

\[
\text{APD} = \int_{X_0}^{f_1} \frac{\tau_f(V^*(f,X))}{f_0(V^*(f,X)) - f} df,
\]

where \( f_0 (X_0) \) and \( f_1 (X_1) \) are the values of the slow variable \( f(X) \) at the start and end points of the depolarized portion on the upper branch of slow manifold in Fig.3, respectively. The integral is evaluated (numerically) along the curve (trajectory) on the upper branch of the slow manifold, and both \( f \) and \( X \) take values along the curve.

3.4. Abnormal (Chaotic) Behavior Induced by Geometrical Conformation Change of Slow Manifold

So far, we have examined the LR model with normal parameter values. However, such parameter values vary by various causes and the variation of parameter values induces serious diseases. Especially, diseases caused by the defect of ionic channel characteristics are called ionic channel diseases. Figure 4 shows an example of abnormal action potential where the value of \([K]_0\) is decreased from the standard (normal) value 5.4 to 1.0. The folded surface
of this figure shows the slow manifold similarly to Fig.3 and a dark messy curve shown near the middle of each panel is the projection of the (eight-dimensional) full system’s trajectory to the f-X-V plane. Two panels (a) and (b) show the slow manifold and full system’s trajectory from two different viewpoints. The geometrical conformation of slow manifold is changed by the decrease of [K]o from the normal case (Fig.3). This conformation change of slow manifold changes the stable equilibrium to the unstable one and the membrane potential oscillates (chaotically) without any current input (Iext = 0). We note that the small mess oscillation of this chaotic orbit locates near the ‘uppernee’ of the slow manifold. This chaotic orbit has a similar structure to a so-called ‘canard’ solution [5]; it moves along not only the upper and lower stable branches but also the middle ‘unstable’ branch of the slow manifold. More detailed analysis of this chaotic oscillation and other examples will be shown at the presentation.

Acknowledgments

This research was partially supported by the Japanese Ministry of Education, Science, Sports and Culture, Grant-in-Aid (#18500172) and by Artificial Intelligence Research Promotion Foundation.

Appendix

Formulæ of the LR model and standard (normal) values of parameters are listed below.

\[ I_{INa} = \text{Na}_4 \cdot \text{j}_{i} \cdot (V - E_{Na}) \]
\[ \text{Na}_5 = 23.0, \quad E_{Na} = 54.4 \]
\[ \alpha_{Na}(V) = 0.32(V - 47.13) \]
\[ \beta_{Na}(V) = 0.08 \exp\left(-\frac{V}{44}\right) \quad (V \geq -40) \]
\[ \alpha_{Na}(V) = 0.0 \quad (V < -40) \]

\[ \beta_{K}(V) = 0.135 \exp\left(-\frac{(V + 80)/6.8}{6.8}\right) \quad (V \geq -40) \]
\[ (1/0.135 \cdot \exp\left(-\frac{(V + 10.65)/11.1}{11.1}\right)) \quad (V \geq -40) \]
\[ 3.56 \exp\left(0.0799V - 3.1 \times 10^{-4} \exp\left(0.35V\right)\right) \quad (V < -40) \]
\[ a_{Na}(V) = \frac{0}{1 + \exp(9.31V + 79.23)} \quad (V < -40) \]
\[ 0.3 \exp(-2.53V^2) \quad (V \geq -40) \]
\[ \beta_{K}(V) = 0.1212 \exp\left(-0.01052V\right) \quad (V \geq -40) \]

References

Predicting Instability of Power Systems based on Hybrid System Reachability Analysis

Yoshihiko Susuki, Takashi Hikihara, and Hiroaki Ebina†

Kyoto University, Department of Electrical Engineering
Katsura, Nishikyo, Kyoto, 615–8510, Japan
Email: susuki@dove.kuee.kyoto-u.ac.jp, hikihara@kuee.kyoto-u.ac.jp

Abstract—This paper shows a numerical approach to predicting instability of a power system based on reachability analysis of hybrid systems. The analysis is performed by computing reachable sets of unsafe sets in hybrid automata that represent both continuous dynamics of generators and discrete operations by relay devices. The unsafe sets here are subsets of state space in which a power system shows unacceptable operations. Then instability of power system can be estimated by investigating whether a system state exists in the reachable sets or not. The estimation is possible at any time in operation, for example, onset of relay operations. This paper numerically demonstrates the proposed approach via analyses of simple power transmission systems.

1. Introduction

There is growing recognition that the dynamics of electric power systems have become complicated as the current change of technological bases and economic environment [1, 2]. Various power apparatuses including HVDC systems and FACTS are applied to conventional ac transmission systems. Regulatory reforms of electricity markets require a substantial modification of conventional system operation. These possibly cause the dynamics of power systems to become complicated. It is by now widely recognized that we cannot fully analyze and control such complex dynamics using conventional power engineering. A comprehensive approach to the analysis and control has been therefore strongly required [3].

This paper introduces a numerical approach to predicting instability of a power system based on hybrid systems theory. The concept of instability here is concerned with a system’s ability to maintain an acceptable operating condition following a load variation or event disturbance [4]. Hybrid dynamical systems and their control are active research subjects in computer science and control engineering [5]. Several researchers have recently worked on the intersection of power system analysis and hybrid systems theory [6, 7, 8, 9, 10, 11, 12, 13, 14]. A general framework for modeling and stability analysis of power systems based on hybrid systems theory is also proposed in [15]. This paper focuses on instability problems and introduces a numerical approach to predicting the instability of power system. A key idea of the approach is to compute backward reachable sets of unsafe sets in hybrid automata as power system models. This paper demonstrates the proposed approach via analyses of simple power transmission systems. The contents in this paper are presented in [16, 17, 18].

2. Predicting Instability of Power System

This section introduces a numerical approach to predicting the instability of power system [15, 16].

2.1. Hybrid Automaton as Power System Instability Model

A hybrid automaton [19] can combine continuous dynamics of power systems with discrete operations and disturbances. The hybrid automaton is a well-known formalism of hybrid systems which comprises continuous vectorfields and discrete transitions. The continuous vector fields can represent electromagnetic and voltage dynamics of generators and dynamic response of generator control systems and transmission equipments. The discrete transitions can model topological changes of transmission networks and changes of transformer tap positions and shunt capacitor states. The hybrid automaton can also include continuous control and disturbances such as dc links, SVCs, and unanticipated power flow, and discrete control and disturbance actions such as relay operations and event faults including lightning and timber contact. The hybrid automaton is hence applicable to the modeling of continuous dynamics in power systems with taking discrete operations into account.

2.2. Predicting Instability Using Reachability Analysis of Hybrid Automaton

This subsection explains a novel approach to predicting the instability of power system. Now define
an unsafe set $G$ [20] in state space of the hybrid automaton. The unsafe set is interpreted as a subset of operating conditions in which a power system shows unacceptable behavior: occurrence of large rotor speed deviation, stepping-out of generators, and very low or high voltages. A reachable set $R_t(G)$ [20] for time $t < 0$ is then defined by a subset of state space from which any trajectory reaches the boundary $\partial G$ of $G$ in $|t|$ time despite of any control. Fig. 1 shows the concept of reachable sets in continuous state space. The concept of reachable sets is important for estimating the instability of power system. If a system state exists in $R_t(G)$, then we can estimate that the power system will reach unacceptable operations in $|t|$ time. The estimation is possible at any time in operation such as onset time of event faults, clearing and re-closing operations. Namely, by computing reachable sets of hybrid automaton, we can consider at any time whether the power system reaches unacceptable conditions as time passes. The reachability analysis thus makes it possible to predict the instability of power system.

3. Application of Simple Power Transmission Systems

This section demonstrates the proposed approach via analyses of simple power transmission systems. The details of modeling and numerical simulations are given in every refereed paper in the following subsection titles.

3.1. Transient Angle Instability [17]

The proposed approach is now applied to the analysis of transient angle instability in single machine-infinite bus (SMIB) system in Fig. 2. The SMIB system consists of a synchronous generator, an infinite bus [21], and two parallel transmission lines. Suppose that the power system undergoes a large fault in one transmission line at time 0 s, that the faulted line is cleared at time $t_c > 0$ s by relay operations, and that the line is re-closed at time $t_{rc} > t_c$. The present application aims to investigate whether the generator remains in synchronism, following the event fault, relay operations, and continuous dynamics of generator.

The interaction of relay operations and continuous dynamics of generator is represented by the hybrid automaton in Fig. 3. $q_i$ for $i = 1, 2, 3$ represents the discrete variable: $q_1$ is assigned to the fault-on operation, $q_2$ the one line operation, and $q_3$ the two lines operation. $\sigma^i$ for $i = 1, 2$ denotes the control action: $\sigma^1$ represents the clearing operation and $\sigma^2$ the re-closing operation. The continuous electro-mechanical dynamics of generator are described by a differential equation as follows:

$$\dot{\delta} = \omega, \quad \dot{\omega} = p_m - b \sin \delta - D \omega. \quad (1)$$

$\delta$ is the rotor position with respect to synchronous reference axis and $\omega$ the rotor speed deviation relative to system angular frequency. $p_m$, $b$, and $D$ are the system parameters. The changes of discrete variables by the control actions $\sigma^i$ are represented by discrete transitions. The unsafe set $G$ here is defined by the subset of state space satisfying $|\omega| \geq 2.0$. Any state in $G$ physically implies the occurrence of large rotor speed deviation and stepping-out.

Figure 4 shows a numerical result of hybrid reachable set under $t_c = 0.1$ s and $t_{rc} = 0.6$ s. The numerical integration is performed in sufficiently large time.
3.2. Transient Voltage Instability [18]

The proposed approach in Section 2 is next applied to the analysis of transient voltage instability in single machine-load bus (SMLB) system with automatic voltage regulator (AVR) in Fig. 5. The SMLB system is given by Venkarasbramanian et al. [22] and consists of a synchronous generator with AVR, a load bus, and ac transmission lines. The application here aims to investigate whether the load bus voltage reaches unacceptable levels, depending on a relay operation and continuous dynamics of synchronous machine. This subsection predicts the behavior of load bus voltage at time 0 s and supposes that one transmission route with two lines is tripped at time $t_c$ (>$0$ s).

The interaction of relay operation and continuous voltage dynamics is formulated by the hybrid automaton in Fig. 6. $q_i$ for $i = 1, 2$ represents the discrete variable: $q_1$ is assigned to the two routes operation and $q_2$ the one route operation. $\sigma$ denotes the control action which represents the relay operation. The continuous voltage dynamics are described by a differential-algebraic equation (DAE) [22] as follows:

$$\dot{x} = f(x, E), \quad 0 = g(x, E; x_1),$$

where $x = (E', E_{id})^T$. $E'$ is generator voltage behind transient reactance, $E_{id}$ the field excitation voltage, and $E$ the load bus voltage. $x_1$ is the transmission reactance and discontinuously changes at time $t_c$. A discrete transition is then defined by discontinuous solutions [23, 24] of the DAE. The unsafe set $G$ is now the subset of state space satisfying $E < 0.7$. Any state in $G$ physically implies the decrease of load bus voltage and the occurrence of voltage instability.

Figure 7 shows a numerical result of hybrid reachable set under $t_c = 2.5$ s. The numerical integration is also performed in sufficiently large time at the discrete state $q_2$. The figure shows the set of initial conditions at time 0 s and the unsafe set $G$. The reachable set is decomposed into the three subsets $R_1$, $R_2$, and $R_3$. $R_1$ is the subset from which any trajectory reaches $\partial G$ at the discrete variable $q_1$. $R_2$ is the subset from which any trajectory directly lands on $G$ by the discrete transition. $R_3$ is also the subset from which any trajectory reaches $\partial G$ after the discrete transition. The region $NR$ is the subset of $I$ in which any discrete transition cannot be defined. $x_S$ in Fig. 7 exists in the complement of reachable sets and it is therefore estimated at 0 s that the power system does not show the unacceptable behavior as time passes. $x_1$ in the figure is in the reachable set $R_1$, and it can be predicted that the power system reaches the unacceptable behavior in time $t_c$. The prediction of transient voltage instability is hence achieved via the computation of hybrid reachable set.

4. Summary

This paper explained a numerical approach to predicting the instability of power system based on hybrid systems theory. The proposed approach is based
on modeling of power system dynamics via hybrid automata and their reachability analysis. The main contribution of this paper is to show that by computation of hybrid reachable sets we can predict the instability of power system at any time in operation.

Acknowledgments
The authors are grateful to Mr. Takeji Uemura, Mr. Takashi Ochi, Mr. Kosaku Yokota, and Mr. Takuya Sakiyama in Kansai Electric Power Co., Ltd. for valuable discussions. This research is partially supported by the Ministry of Education, Culture, Sports, Sciences and Technology in Japan, The 21st Century COE Program #14213201, and Young Research (B) #18760216, Innovation in Japan, The 21st Century COE Program

References

Fuzzy measures and integrals in re-identification problems

Jordi Nin and Vicenc Torra

Institut d’Investigació en Intel·ligència Artificial - CSIC
Campus UAB s/n
08193 Bellaterra (Catalonia, Spain)
Email: vtorra@iiia.csic.es, jnin@iiia.csic.es

Abstract—In this paper we give an overview of our approach of using aggregation operators, and more specifically, fuzzy integrals for solving re-identification problems. We show that the use of Choquet integrals are suitable for some kind of problems.

Keywords: Record linkage, Fuzzy integrals, OWA operators, data mining, data cleaning, privacy preserving data mining.

1. Introduction

Re-identification algorithms when applied to databases permit to identify those objects that can be found in different files but that correspond to the same entity. Two particular family of algorithms can be distinguished.

- **Record linkage (or record matching) algorithms:** These algorithms intend to link records of one file with those records in another file that correspond to the same individual. The difficulties of the approach are due to the fact that the records might be described using different attributes, or, in the case of using the same attributes, there are errors in the data.

- **Schema matching or, more particularly, attribute matching:** They are algorithms to link attributes or schemata in databases. The typical problems these algorithms have to face is that the name of the attributes in different files do not coincide, or that a single attribute in one file corresponds to several ones in another file. In the most general case, schema matching needs to construct $n:m$ relationships.

In recent years we have shown that some particular fuzzy integrals can be used in some cases for re-identification. They have been applied to both record linkage and attribute matching.

In this paper we will describe how this problem is tackled using fuzzy integrals, the underlying assumptions of our model, and describe why this approach works. Some simple examples of application will be given.

The structure of the paper is as follows. In Section 2 we give some preliminaries. We focus on fuzzy measures and integrals and on the re-identification methods. Then, in Example 3 we present an example. The paper finishes with some conclusions.

2. Preliminaries

This section is devoted to give a short review of fuzzy measures and integrals and then, to some re-identification methods.

2.1. Fuzzy measures and integrals

**Definition 1** A fuzzy measure $\mu$ on a set $X$ is a function $\mu : 2^X \rightarrow \mathbb{R}^+$ with the following properties:

1. $\mu(\emptyset) = 0$
2. $m(A) \leq m(B)$ whenever $A \subseteq B$ and $A, B \in 2^X$.

For the sake of simplicity, we assume $X = \{1, \ldots, N\}$. Families of fuzzy measures have been defined in the literature. In this paper we focus on the so-called symmetric fuzzy measures. In a symmetric fuzzy measure, the measure of a set only depends on the cardinality of the set but not on the elements of the set.

**Definition 2** A fuzzy measure $\mu$ is symmetric when $\mu(A) = \mu(B)$ whenever $|A| = |B|$.

Here, $|\cdot|$ represents the cardinality of a set.

Symmetric fuzzy measures can be represented by non-decreasing functions $f : [0, 1] \rightarrow [0, 1]$ such that $f(0) = 0$ and $f(1) = 1$ so $\mu(A) = f(|A|/|X|)$.

**Definition 3** [1] Given a fuzzy measure $\mu$ and a function $f$, the Choquet integral of $f$ with respect to $\mu$ is defined using the i-th order statistics $(x_{(i)})$ as:

$$CI_\mu(f) := \sum_{i=1}^{n} f((x_{(i)}) - f(x_{(i-1)}))\mu((i) \cdot (n))$$

where we define $x^{(0)} := 0$.

The Choquet integral with respect to a symmetric fuzzy measure corresponds to the OWA operator [9].

**Definition 4** [4] Given a fuzzy measure $\mu$ and a function $f$ into $[0, 1]$, the Sugeno integral of $f$ with respect to $\mu$ is defined using the i-th order statistics $(x_{(i)})$ as:

$$S_\mu(f) := \bigvee_{j=1}^{j=n} f(x_{(i)}) \wedge \mu(A_{(i)})$$

$A_{(i)} = \{x_{(i)}, \ldots, x_{(n)}\}, A_{(m+1)} = \emptyset$. Here $\wedge$ denotes the minimum and $\vee$ denotes the maximum.
2.2. Re-identification methods

As said in the introduction, standard record linkage methods are centered on the linkage of objects belonging to the same entities from two or more files when such files share a set of variables, or any other kind of information. In this case, the difficulties for a good performance of record linkage algorithms are due to the fact that files contain errors (e.g., the income of an individual is not the same in both files or attributes are represented in different scales).

Two main approaches have been used for standard record linkage. See [6, 7, 8] for more details:

Probabilistic Record Linkage(PRLB): For each pair of records \((a, b)\), we compute a conditional probability of having a correct link using a coincidence vector of variables. Then, we use this probability to classify each pair \((a, b)\) as either a linked pair (LP) or a non-linked pair (NP).

Distance-based Record Linkage(DBRL): Records of file \(A\) are compared to records of file \(B\) with respect to a given distance function, and then each record in \(A\) is linked to the nearest record in \(B\) using such distance function.

3. Example: normal distributions

We describe below some experiments to test the performance of fuzzy integrals in a re-identification problem. The fuzzy integrals have been combined with fuzzy measures.

The experiments use normal distributions as entities, so a normal distribution is considered as an individual or attribute (either record linkage or attribute linkage).

The fuzzy integrals have been computed with respect to symmetric fuzzy measures generated from the following three parametric functions:

\[ Q_{21}(x) = x^\alpha \text{ for } \alpha = 1/5, 2/5, \ldots, 10/5 \]
\[ Q_{22}(x) = 1/(1 + e^{(\alpha-3)10}) \text{ for } \alpha = [0, 0.1, \ldots, 0.9] \]
\[ Q_{23}(x) = \begin{cases} 0 & \text{if } x \leq \alpha \\ 1 & \text{if } x > \alpha \end{cases} \text{ for } \alpha = [0, 0.1, \ldots, 0.9] \]

3.1. Data generation

We have generated two different synthetic data files using a pseudo-random gaussian generator. The files were generated with twenty normal distribution (entities). The first file contains normal distributions with \((\mu, \sigma)\) from the sets \(\mu = [0, 1, 2, 3, 4]\) and \(\sigma = [0.5, 1, 1.5, 2]\). The second file contains normal distributions from \((\mu, \sigma)\) with \(\mu\) in \(\mu = [0, 2, 4, 6, 8]\) and \(\sigma\) in the same set \(\sigma = [0.5, 1, 1.5, 2]\). For each distribution we have generated twenty-five elements (variables).

These files correspond to the original data files. These distributions are represented in Figure 1.

Afterwards, we have generated four synthetic data files for each original data file. These files have been generated with the same normal distributions but adding linear noise based on a normal distribution \(N(0, 1)\). We have considered four levels of noise, \(\epsilon = 0.1 \ast N(0, 1), \epsilon = 0.25 \ast N(0, 1), \epsilon = 0.5 \ast N(0, 1)\) and \(\epsilon = N(0, 1)\). These files are to be used in the re-identification experiment. These distributions are displayed in Figure 2.

According to this description, we obtain two original files with twenty records (normal distributions) described with twenty-five attributes. Additionally, we get eight distorted files with an identical number of records belonging to the same normal distributions but modified with some noise. The number of attributes is the same, although in this case, they are noisy attributes.

3.2. Re-identification

We have used the fuzzy integrals for re-identification. In particular, we have used the Choquet integral with symmetric fuzzy measures. As said above, such operator is equivalent to the OWA operator.

The integral, with a particular measure, is applied to each record, obtaining a representative for such record. Different parameterizations lead to different representatives.

Then, once we have representatives for all records, re-identification is done comparing the representatives of each record in one file with the representatives of each record in the other file. The rationale is that when the representatives are similar, the original records should also be similar.

In our case, we have considered the fuzzy measures listed above with 10 different parameterizations. This leads
to 10 representatives for each record (for each normal distribution, either original or distorted).

Taking this into account, new files are created with the representatives. That is, for each pair of files, we obtain a second pair of files. These new files have the same number of records as the original files (representatives are built for each original record) and 10 attributes (there is one attribute for each parameterization). As the representatives of both files have been obtained from the same parameterizations, we can say that the new files are described using the same attributes.

Therefore, as both file share attributes (the parameterizations), we can use now standard record linkage methods (e.g., probability or distance-based) to link the files.

3.3. Results

We have obtained good results with all fuzzy measures. The results are given in Table 1. The table contains the best number of re-identifications obtained for each experiment with either distance-based or probabilistic record linkage.

The experiments show that the re-identification is possible, and that the larger the distortion, the worse the re-identification. These are expected results.

<table>
<thead>
<tr>
<th>ε</th>
<th>µ</th>
<th>Case 1</th>
<th>Case 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>N(0,1)*0.1</td>
<td>Q^1_1(x)</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>N(0,1)*0.25</td>
<td>Q^1_1(x)</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>N(0,1)*0.5</td>
<td>Q^1_1(x)</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>N(0,1)*1.0</td>
<td>Q^1_1(x)</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>N(0,1)*0.1</td>
<td>Q^2_1(x)</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td>N(0,1)*0.25</td>
<td>Q^2_1(x)</td>
<td>14</td>
<td>17</td>
</tr>
<tr>
<td>N(0,1)*0.5</td>
<td>Q^2_1(x)</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>N(0,1)*1.0</td>
<td>Q^2_1(x)</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>N(0,1)*0.1</td>
<td>Q^3_1(x)</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>N(0,1)*0.25</td>
<td>Q^3_1(x)</td>
<td>11</td>
<td>18</td>
</tr>
<tr>
<td>N(0,1)*0.5</td>
<td>Q^3_1(x)</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>N(0,1)*1.0</td>
<td>Q^3_1(x)</td>
<td>11</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 1: Results of the re-identification. Case 1: Files with the following normal distributions N(0,0.5)..N(4,2); Case 2: Files with the following normal distributions N(0,0.5)..N(8,2)

3.4. Alternative example

In the previous section we have described the results when the masked file are obtained with linear noise addition. But we can considered other ways to obtain a masked file.

We can consider the addition of other kind of noise like N(μ,σ)+ε*(1−ε)*N(μ’,σ’). In this case we obtain a new probability distribution like the one we can see in Figure 3.

If we apply the Choquet integral with a symmetric measure (OWA operator) to this family of distribution for small $\epsilon$ we will obtain similar results than when the distortion corresponds to the addition of linear noise. This is so because the OWA operator sorts data values from the smallest values to greatest ones and the resulting values after the ordering are not so dissimilar. For larger values of $\epsilon$ this would not be the case, the noisy distribution might be re-identified with a time-series generated from N(μ’,σ’).

4. Conclusions

In this paper, we have studied the use of owa operators for the re-identification individual problem, and focused in the particular case in which individuals are represented by normal distributions. We have proved that owa operators are a suitable tools for such re-identification as they have lead to good results with three different fuzzy measures.

Additional experiments have been done with real data. Some of the results are reported in [5] (attribute matching) and [2, 3] (record linkage).

Acknowledgments

This work was partly funded by the Spanish Ministry of Education and Science under project SEG2004-04352-C04-02 “PROPIETAS”.

References


Generalizing Choquet Integral with Level Dependent Capacities

Salvatore Greco†, Silvio Giove‡, and Benedetto Matarazzo†

† Faculty of Economics, University of Catania
Corso Italia 55, Catania 95129, Italy
‡Faculty of Economics, University of Venice
Dorsoduro 3825/E, Venezia 30123, Italy
Email: salgreco@unict.it, sgiove@unive.it, matarazz@unict.it

Abstract. We present a generalization of Choquet integral in which the capacity depends on the level of the aggregated variables. We show that as particular cases of our generalization of Choquet integral there are the Sugeno integral, the Sipos integral and the Cumulative Prospect Theory functional. We show also that many concepts such as Mobius transform, importance index, interaction index, k-order capacities and OWA operators, introduced in the research about Choquet integral can be generalized in the considered context.

1. Introduction

In many decision problems a set of actions are evaluated with respect to a set of points of view called criteria. For example, in evaluating a car one can consider criteria such as maximum speed, price, acceleration, fuel consumption. In evaluating a set of students one can consider as criteria the evaluations in examinations with respect to different subjects such as mathematics, physics, literature and so on. In general, evaluations with respect to different criteria can be discordant. For example, very often when a car has a good maximum speed, it has also a high price and a high fuel consumption or if a student is very good in mathematics, may be not so good in literature. Thus, in order to express a decision such as a choice among a set of car or a ranking of a set of students, it is necessary to aggregate the evaluations on considered criteria. This is the domain of multiple criteria decision analysis and in this context several methodologies have been proposed (for a collection of extensive state-of-art surveys see [3]).

One of the simplest aggregation procedures is the weighted sum of the evaluation of criteria. Some more complex aggregation procedures have been proposed to take into account specific aspects in evaluating importance of criteria, such as interaction between criteria.

The interaction of criteria has been considered through non-additive integrals such as Choquet integral [2] and Sugeno integral [13] (for a comprehensive survey see [5]). In this context, the importance of a set of criteria is not necessarily the sum of the importance of each criterion in the set. It can be greater or smaller, due to redundancy or synergy between criteria. An example of redundancy is the case of maximum speed and acceleration in evaluating cars: in fact a car that is speed very often has also a good acceleration and therefore, even if these two criteria can be very important for a person liking sport cars, the importance of them is smaller than the importance of the two criteria considered separately. An example of synergy is the case of maximum speed and price in evaluating cars: in fact a car that is speed, is often also a highly priced, and therefore a car with a high maximum speed and a not so high price is very well appreciated. So, the importance of these two criteria is greater than the importance of the two criteria considered separately.

Introduction of a neutral point and consequent positive and negative sign of the evaluations with respect to considered criteria has been taken into account through some generalizations of the Choquet integral such as Sipos integral [12] and Cumulative Prospect Theory functional [14]. In this paper we propose a further generalization of the Choquet integral which takes into account the fact that the importance of criteria depends on the level of their evaluations. For example, in evaluating a car with respect to price, comfort and fuel consumption, for cars with low levels of price, comfort and fuel consumption, the price is more important than the comfort, while for cars having higher levels of price, comfort and fuel consumption, the comfort becomes more important than the fuel consumption. To model representation of the preferences in these situations we propose to assign importance to criteria depending from the level of evaluations and a consequent generalization of the Choquet integral. This is a very flexible model to aggregate evaluations of set of criteria. Moreover, in a very surprising way, this model presents as specific cases the Sugeno integral, the Sipos integral and the Cumulative Prospect Theory functional.

The plan of the paper is as follows. In section 2 we introduce the generalized Choquet integral and characterize it on the basis of some simple and intuitive axioms. In section 3 we show that the classical Choquet integral and the Sugeno integral are specific cases of our generalized Choquet integral. In section 4 we generalize to the Choquet integral the concepts of Mobius decomposition, importance index and interaction indices. Moreover, in the same section, using the generalized Choquet integral, we generalize k-order capacity and Ordered Weighted Aggregation (OWA). In section 8 we prove that Sipos integral and the Cumulative Prospect Theory functional can be interpreted as special cases of...
generalized Choquet integral. The last section presents final conclusions.

2. The generalized Choquet integral

Let us consider a set of criteria $N=\{1,\ldots,n\}$. We define a generalized capacity as a function $\mu^G:2^N\times\{0,1\}\rightarrow\{0,1\}$ such that

1. for all $t\in\{0,1\}$ and $A\subseteq B\subseteq N$, $\mu^G(A,t)\leq\mu^G(B,t)$
2. for all $t\in\{0,1\}$, $\mu^G(\emptyset,t)=0$ and $\mu^G(N,t)=1$.

We consider also the following regularity property:

3. for all $x=[x_1,\ldots,x_n]$, $x\in[0,1]^n$, and for all $t\in\{0,1\}$, the function $f:[0,1]^n\times\{0,1\}\rightarrow[0,1]$ defined as $f(x,t)=\mu^G(A(x,t),t)$, with $A(x,t)=\{i\in N: x_i \geq t\}$, is continuous with respect to $t$ almost everywhere.

We define the generalized Choquet integral of $x=[x_1,\ldots,x_n]$, $x\in[0,1]^n$, with respect to the generalized capacity $\mu$ as follows:

$$Ch(x, \mu^G) = \int_0^1 \mu^G(A,x,t)dt.$$ 

Let us remark that condition 3) ensures that the generalized Choquet integral exists always.

A similar approach has been proposed in the field of decision under uncertainty [1,6,7,10].

Any function $G:[0,1]^n\rightarrow[0,1]$ is an aggregation function. The following properties of an aggregation function $G$ are useful to characterize the generalized Choquet integral:

- **monotonicity:** for all $x,y\in[0,1]^n$, $x\succeq y \Rightarrow G(x)\geq G(y)$,
- **idempotency:** for all $a\in[0,1]^n$ such that $a=[a,\ldots,a]$, $G(a)=a$,
- **tail independence:** for all $x,y,w,z\in[0,1]^n$ and $A\subseteq N$ such that, for all $i\in A$ and $j\in N-A$

$$x_i\geq y_i, x_{2i}z_i, y_{i2}w_i, y_{ij}z_j$$

we have $G(x,y,w,z)=G(y,w,z)=G(y,z)=G(y,z)$ where, for all $h,k\in[0,1]^n$, $m=[h,k]$ is defined in such a way that, for all $i\in N$, $m_i=h_i$ if $i\in A$ and $m_i=k_i$ if $i\notin A$.

**Theorem 1.** An aggregation function $G:[0,1]^n\rightarrow[0,1]$ is monotonic, idempotent and tail independent if and only if there exists a generalized capacity $\mu^G$ such that, for all $x\in[0,1]^n$

$$G(x)=Ch(x, \mu^G).$$

3. The generalized Choquet integral, the classical Choquet integral and the Sugeno integral

**Definition.** A generalized capacity $\mu^G$ is level independent if $\mu^G(A,t)=\mu^G(A,u)$ for all $t,u\in\{0,1\}$.

**Remark 1.** If the generalized capacity $\mu^G$ is level independent, then the generalized Choquet integral collapses to the standard Choquet integral [2].

**Definition.** A generalized capacity $\mu^G$ is Boolean if $\mu^G(A,t)\in\{0,1\}$ for all $A\subseteq N$ and for all $t\in\{0,1\}$.

**Theorem 2.** Given a generalized capacity $\mu^G$, there exists another generalized capacity $\mu^G(A,t)$ Boolean and monotonic non increasing with respect to $t$, such that for all $A\subseteq N$, $\mu^G(A,t)$ almost everywhere with respect to $t\in\{0,1\}$ if and only if there exists a capacity $v$ on $N$ (i.e. a function $v:[0,1]^{2^N}\rightarrow[0,1]$) for which $v(A)\leq v(B)$ for all $A\subseteq B\subseteq N$, $v(\emptyset)=0$, $v(N)=1$ such that, for all $x\in[0,1]^n$,

$$Ch(x, \mu^G)=\max\left(\min\{\min\{x_i, i\in A\}, v(A)\}\right),$$

i.e. the generalized Choquet integral is the Sugeno integral of $x$ with respect to the capacity $v$.

Remark 1 and Theorem 2 say that the generalized Choquet integral contains as particular cases the Choquet integral and the Sugeno integral. It is interesting to observe, that, anyway, it is much more general.

4. Mobius Transform, importance indices and interaction indices

4.1. Mobius transform of a generalized capacity.

With respect to classical Choquet integral, Mobius transform has been deeply investigated in [4].

Given a generalized capacity $\mu^G$ the Mobius transform of a generalized capacity is a function $m^G:2^N\rightarrow\{0,1\}$ such that, for all $A\subseteq N$ and all $t\in\{0,1\}$, we have that $\mu^G(A,t)=\sum_{B\subseteq A} m^G(B,t)$.

Expliciting Mobius transform we obtain that for all $A\subseteq N$ and all $t\in\{0,1\}$

$$m^G(A,t)=\sum_{B\subseteq A} (-1)^{|A\setminus B|} \mu(B,t).$$

**Theorem 3.** For all $x\in[0,1]^n$ and for all generalized capacity $\mu^G$

$$Ch(x, \mu^G)=\sum_{B\subseteq N} m^G(B,t)dt.$$
The Shapley value \( \phi(\mu^G, i, t) \) is an average with “proper weights” \( \mu^G(A \cup \{i\}, t) - \mu^G(A, t) \) of marginal contribution \( \mu^G(A \cup \{i\}, t) \) of criterion \( i \) to all subsets \( A \) from \( N \). It measures the importance of criterion \( i \in N \) with respect to generalized capacity \( \mu^G \) and level \( t \in [0,1] \).

Importance index \( \phi(\mu^G, i, t) \) can be represented in terms of Mobius transform as follows:

\[
\phi(\mu^G, i, t) = \sum_{A \subseteq N} \frac{m^G(A \cup \{i\}, t)}{|A|} \mu^G(A, t)
\]

An overall importance index \( \phi(\mu^G, i) \) averaging the importance indices \( \phi(\mu^G, i, t) \) relative to all level \( t \), \( t \in [0,1] \), can be calculated as follows:

\[
\phi(\mu^G, i) = \frac{1}{|N|} \int_0^1 \phi(\mu^G, i, t) dt.
\]

### 4.3. Interaction indices

With respect to classical Choquet integral, interaction indices have introduced in [9] with respect to only pairs of criteria and in [4] with respect to all possible subsets of criteria.

Given a generalized capacity \( \mu^G, A \subseteq N \) and \( t \in [0,1] \) the Shapley interaction index can be calculated as:

\[
I(\mu^G, A, t) = \sum_{B \subseteq N-A} \frac{|B|}{|N|} \left( \sum_{C \subseteq A} (-1)^{|C|-1} \mu^G(C \cup B, t) \right)
\]

Interaction index \( I(\mu^G, A, t) \) can be represented in terms of Mobius transform as follows:

\[
I(\mu^G, A, t) = \sum_{B \subseteq N-A} \frac{m^G(A \cup \{i\}, t)}{A \cup B} \mu^G(A, t).
\]

The Shapley interaction index \( I(\mu^G, i, t) \) is an average with “proper weights” \( \frac{|B|}{|N|} \left( \sum_{C \subseteq A} (-1)^{|C|-1} \mu^G(C \cup B, t) \right) \) of interaction between criteria from \( A \) when adjoined to all subsets \( B \) from \( N-A \). It measures the interaction of criteria from \( A \) with respect to the generalized capacity \( \mu^G \) and the level \( t \in [0,1] \).

An overall interaction index \( I(\mu^G, A, t) \) averages the interaction indices \( I(\mu^G, A, t) \) relative to all levels \( t \), \( t \in [0,1] \), then it can be calculated as follows:

\[
I(\mu^G, A, t) = \frac{1}{|A|} \int_0^1 I(\mu^G, A, t) dt.
\]

### 4.3. k-order generalized capacity

With respect to the classical Choquet integral, the k-order capacity has been introduced in [4].
Theorem 5. If the generalized capacity is bi-partite, then
\[ Ch(\mathbf{x}, \mu^G) = \int_0^1 \mu^+_{G} \left( \{ i : x_i - e \geq t \} \right) dt + \]
\[ - \int_0^e \mu^-_{G} \left( \{ i : e - x_i \geq t \} \right) dt + e \]

where, for all \( A \subseteq N \), \( \mu^+_{G}(A) = \mu^+(A) \) and \( \mu^-_{G}(A) = 1 - \mu^-(N - A) \).

Let us remark that according to Theorem 5, we can write
\[ Ch(\mathbf{y}, \mu) = Ch(x, \mathbf{e}) + Ch(\mathbf{x} - \mathbf{e}, \mu^-_{G}) + Ch(\mathbf{x} - \mathbf{e}, \mu^+_{G}) + e \]

where \( Ch(\mathbf{y}, \mu) \) denotes the classical Choquet integral of \( y \in [0,1]^n \) with respect to capacity \( \mu \), \( x \cdot \mathbf{e} \) is the vector having as i-th element \( x_i - e \), \( y^+ \) is the positive part of vector \( y \in [0,1]^n \) having as i-th element \( y_i \), if \( y_i \geq 0 \) and 0 if \( y_i < 0 \), \( y^- \) is the negative part of vector \( y \in [0,1]^n \) having as i-th element \( -y_i \), if \( y_i \leq 0 \) and 0 if \( y_i > 0 \). Let us remember that
\[ Ch(y), \mu^+_{G} + Ch(y), \mu^-_{G} \]
is the Cumulative Prospect Theory functional [14] of \( y \in [0,1]^n \) with respect to the pair of capacities \(( \mu^+_{G}, \mu^-_{G})\). This theorem says that even the Cumulative Prospect Theory functional can be written in terms of generalized Choquet integral. A specific case of the Cumulative Prospect Theory functional is given by the case in which \( \mu^+_{G}(A) = \mu^-_{G}(A) \) for all \( A \subseteq N \). This specific case constitutes the Sipos integral [12] that, therefore, can be also expressed in terms of the generalized Choquet integral.

6. Conclusions

We presented a generalization of Choquet integral in which the capacity depends on the level of the aggregated variables. We showed that as particular cases of our generalization of Choquet integral there are the Sugeno integral, the Cumulative Prospect Theory functional and the Sipos integral. We showed also that many concepts such as Mobius transform, importance index, interaction index, k-order capacities and OWA, can be generalized in the considered context.

Acknowledgments

The research has been supported by the Italian Ministry of Education, University and Scientific Research (MIUR).

References

Abstract—Rough set analysis is often applied to multi-attribute decision problems. Given a decision table, attribute importance has been evaluated in three ways through cooperative games constructed by rough set analysis. In this paper, we extend the approach considering a case when a decision table may include errors. We show that attribute importance can be evaluated from various viewpoints in the case. It is emphasized that attribute importance should be investigated from several perspectives.

1. Introduction

Rough set analysis [1] is one of useful tools to analyze multi-attribute decision problems. The important attributes can be known from the core and reducts and decision rules can be induced from the rough set analysis of decision tables. Moreover, attribute importance is evaluated numerically by cooperative games constructed from a decision table through rough set analysis.

Given a decision table, quality of approximation and classification power index [2] are defined based on the classical rough set model. Those indices are closely related to reduct preserving all lower approximations (for short, L-reduct) and reduct preserving all upper approximations (for short, U-reduct), respectively. Constructing a cooperative game corresponding to each index, we can define a degree of attribute importance by the Shapley value or the Banzaf game corresponding to each index, we can define a degree of attribute importance by the Shapley value or the Banzaf game. We demonstrate the variety of attribute importance.

In this paper, we propose a way to construct cooperative games corresponding to the structure-enhancing reducts proposed by Inuiguchi [6]. The attribute importance can be evaluated by Shapley values of the constructed cooperative games. We demonstrate the variety of attribute importance. We show that the proposed approaches are extensions of previous approaches.

This paper is organized as follows. In Section 2, we briefly introduce a variable precision rough set model and structure-enhancing reducts. A way to construct a cooperative game with respect to a given kind of reduct is proposed in Section 3. Moreover, in Section 3, it is shown that the previous attribute importance defined in the classical rough set model is obtained as a special case. In Section 4, the proposed approach is applied to a didactic example so that the variety in evaluations of attribute importance is demonstrated.

2. Variable Precision Rough Sets and Reducts

2.1. Decision Table

A decision table is characterized by 4-tuple $\mathcal{I} = \langle U, C \cup \{d\}, V, \rho \rangle$, where $U$ is a finite set of objects, $C$ is a finite set of condition attributes, $d$ is a decision attribute, $V = \bigcup_{a \in C \cup \{d\}} V_a$ and $V_a$ is a domain of the attribute $a$, and $\rho: U \times C \cup \{d\} \rightarrow V$ is an information function such that $\rho(x, a) \in V_a$ for every $a \in C \cup \{d\}, x \in U$.

Given a set of attributes $A \subseteq C \cup \{d\}$, we can define an equivalence relation $I_A$ referred to as an indiscernibility relation by

$I_A = \{(x, y) \in U \times U \mid \rho(x, a) = \rho(y, a), \forall a \in A\}$.

From $I_A$, we have equivalence classes,

$[x]_A = \{y \in U \mid (y, x) \in I_A\}, x \in U.$
A shows a set of objects which cannot be discerned from $x$ by information about all attributes in $A$. When $A = \{d\}$, we define

$$
D = \{D_j, j = 1, \ldots, p\} = \{[x]_A \mid x \in U\},
$$

where we have $D_i \neq D_j$ ($i \neq j$) and $D_j$ is called a decision class. There exists some $v_j \in \hat{V}_d$ such that $ho(x, d) = v_j$ for each $x \in D_j$.

### 2.2. The Variable Precision Rough Set Model

Let $A \subseteq C$. For $x \in U$, a rough membership value to a decision class $D_j$ is defined by

$$
\mu_{D_j}(x|A) = \frac{|[x]_A \cap D_j|}{|[x]_A|},
$$

where $|X|$ shows the cardinality of a set $X$. $\mu_{D_j}(x|A)$ shows the ratio of objects in a decision class $D_j$ among objects indiscernible from $x$ by information about attributes in $A$. If $\mu_{D_j}(x|A) = 1$, we have $[x]_A \subseteq D_j$. On the other hand, if $\mu_{D_j}(x|A) = 0$, we have $[x]_A \cap D_j = \emptyset$. Therefore, $\mu_{D_j}(x|A)$ can be interpreted as a degree of inclusion $[x]_A \subseteq D_j$ holds or a degree of implication "$y \in [x]_A$ implies $y \in D_j$".

In other words, if we cannot discern $y \in [x]_A$ from $x$, then $y$ and $x$ are regarded as equivalent. The equivalent objects are assumed to be in the same decision class. $\mu_{D_j}(x|A)$ shows what extent this assumption is correct.

Given an allowable error ratio $\beta \in [0, 0.5)$, in the variable precision rough set (VPRS) model, lower and upper approximations of $X \subseteq U$ are defined as follows:

$$
\underline{A}_\beta(D_j) = \{x \in U \mid \mu_{D_j}(x|A) \geq 1 - \beta\},
$$

$$
\overline{A}_\beta(D_j) = \{x \in U \mid \mu_{D_j}(x|A) > \beta\}.
$$

A pair $(\underline{A}_\beta(D_j), \overline{A}_\beta(D_j))$ is called a variable precision rough set of $D_j$ with error ratio $\beta$. The boundary region of $D_j$ with error ratio $\beta$ is defined by $BN^\beta_A(D_j) = \overline{A}_\beta(D_j) - \underline{A}_\beta(D_j)$. When $\beta = 0$, the VPRS model of $D_j$ with error ratio $\beta$ is reduced to the classical RS model. $\underline{A}_\beta(D_j)$ is a collection of objects belonging to $D_j$ with high probability $1 - \beta$. On the other hand, $\overline{A}_\beta(D_j)$ is a collection of objects with probability more than $\beta$. We call $\underline{A}_\beta(D_j), \overline{A}_\beta(D_j)$ and $BN^\beta_A(D_j)$ a $\beta$-lower approximation of $D_j$, a $\beta$-upper approximation of $D_j$ and a $\beta$-boundary region, respectively.

When $1/p > \beta$, $UNP^\beta_A(D) = U - \bigcup_{j=1,\ldots,p} \overline{A}_\beta(D_j)$ is not always empty. Then $UNP^\beta_A(D)$ is called the $\beta$-unpredictable region of a partition $D$, hence, given a decision table, we obtain $\beta$-lower approximations $\underline{A}_\beta(D_j), j = 1, \ldots, p$, $\beta$-upper approximations $\overline{A}_\beta(D_j), j = 1, \ldots, p$, $\beta$-boundary regions $BN^\beta_A(D_j), j = 1, \ldots, p$ and $\beta$-unpredictable region $UNP^\beta_A(D)$ as classification structures. Among those, $\beta$-lower approximations $\underline{A}_\beta(D_j), j = 1, \ldots, p$, $\beta$-boundary regions $BN^\beta_A(D_j), j = 1, \ldots, p$ and $UNP^\beta_A(D)$ compose a partition.

### 2.3. Reducts in the VPRS Model

Superfluous attribute information is often included in a decision table. For removing the superfluous information, L-reducts [2] and U-reducts [2] are proposed. An L-reduct (resp. U-reduct) is a minimal set of condition attributes preserving lower approximations (resp. upper approximations) where lower approximations (resp. upper approximations) are regarded as essential information contained in the decision table.

The concept of reduct can be inherited in the VPRS model. The $\beta$-reduct is proposed as a reduct in the VPRS model, which is a minimal set of condition attributes preserving the quality of classification for $A \subseteq C$ defined by

$$
\gamma^\beta_A(D) = \frac{\sum_{j=1}^{p} |\underline{A}_\beta(D_j)|}{|U|}.
$$

Here we note that the minimality is defined over all its subsets in the VPRS model. Moreover, $\beta$-reduct is equivalent to L-reduct when $\beta = 0$.

However, $\underline{A}_\beta(D_i) = \underline{C}_\beta(D_i), i = 1, \ldots, p$, is not always satisfied with $\beta$-reduct $A \subseteq C$. Neither $\underline{A}_\beta(D_i) \subseteq \underline{C}_\beta(D_i), i = 1, \ldots, p$ is. Namely, it may destroy $\beta$-lower approximations. This is because the $\beta$-lower approximations are not monotone with respect to $\beta$. This is not an appropriate property since $\beta$-lower approximations are regarded as the essential information.

Inouiguchi [5, 6] proposed structure-preserving reducts and structure-enhancing reducts. Structure preservation is a strong condition and a structure-preserving reduct does not reduce condition attributes very much. We consider structure-enhancing reducts in this paper.

As structure enhancing properties, the following five basic requirements for a reduct $A \subseteq C$ are considered:

- **L$^\beta$-reduct**: $A \subseteq C$ is called an L$^\beta$-reduct if and only if it is a minimal set satisfying $(L)$.
- **U$^\beta$-reduct**: $A \subseteq C$ is called a U$^\beta$-reduct if and only if it is a minimal set satisfying $(U)$.
- **UN$^\beta$-reduct**: $A \subseteq C$ is called a UN$^\beta$-reduct if and only if it is a minimal set satisfying $(UN)$.
- **LUN$^\beta$-reduct**: $A \subseteq C$ is called an LUN$^\beta$-reduct if and only if it is a minimal set satisfying $(L)$ and $(U)$.
- **UUN$^\beta$-reduct**: $A \subseteq C$ is called a UUN$^\beta$-reduct if and only if it is a minimal set satisfying $(U)$ and $(UN)$.
- **UB$^\beta$-reduct**: $A \subseteq C$ is called a UB$^\beta$-reduct if and only if it is a minimal set satisfying $(U)$ and $(UB)$.
• ˇB ˇUN β-reduct: A ⊆ C is called a ˇB ˇUN β-reduct if and only if it is a minimal set satisfying (B) and (ˇUN).
• LˇUN β-reduct: A ⊆ C is called an LˇUN β-reduct if and only if it is a minimal set satisfying (L), (ˇU) and (ˇUN).
• LˇB ˇUN β-reduct: A ⊆ C is called an LˇB ˇUN β-reduct if and only if it is a minimal set satisfying (L), (ˇU) and (ˇUN).

The strong-weak relation among those 10 reducts is depicted in Figure 1. In Fig. 1, names of reducts are abbreviated to their first characters. Reducts located in the upper part of the figure enhance the structure more than those located in the lower part. Namely a ˇUN β-reducts is the least enhancing the structure. On the other hand, an LˇUN β-reduct and a LˇB ˇUN β-reduct are the most enhancing the structure.

3. Attribute Importance in Decision Tables

3.1. Under Classical Rough Sets

The importance of each condition attribute can be evaluated as the Shapley value or the Banzaf value of a cooperative game defined in a decision table.

Corresponding to L-reduct, a game v L : 2 C → [0, 1] is defined by

\[ v_L(A) = \gamma_A(D) = \frac{\sum_{i=1}^{p} |A_i(D_i)|}{|U|}, \quad \forall A \subseteq C, \]

where we define \( \vartheta_+(X) = \emptyset \) for any \( X \subseteq U \). \( \gamma_A \) is a quality of approximation [2] and \( \gamma_A \) is an extension of \( \gamma_A \).

Moreover, corresponding to U-reduct, \( \partial(x) = \{ D_i \mid x \in \overline{C}(D_i), \ i = 1, \ldots, p \} \) is defined. Using \( \partial \), we can define an equivalence relation by \( I_{\partial} = \{ (x, y) \mid \partial(x) = \partial(y), \ x, y \in U \} \). Accordingly, we have a partition \( U/I_{\partial} = \{ E_1, \ldots, E_q \} \), where \( E_i \) is an elementary set of \( U/I_{\partial} \). Finally a game \( v_U : 2^C \rightarrow [0, 1] \) can be defined by

\[ v_U(A) = \pi_A(D) = \gamma_A(U/I_{\partial}) = \frac{\sum_{i=1}^{q} |A_i(E_i)|}{|U|}, \quad \forall A \subseteq C, \]

\( \pi_A(D) \) is called the classification power index. Note that \( v_U(A) \) takes the maximum value 1 if and only if \( A \) satisfies \( A_0(D_i) = \overline{C}_0(D_i), \ i = 1, \ldots, p \).

Once a cooperative game \( v \) is defined, we can evaluate attribute importance by the Shapley value or the Banzaf value. In this paper, we use the Shapley value, defined by

\[
\phi(a : v) = \sum_{A \subseteq U - \{a\}} \frac{|U| - |A| - 1|A|!}{|U|!} \{ v(A) - v(A - \{a\}) \}.
\]

Extending the Shapley value and the Banzaf value, we can also evaluate the interaction among condition attributes [2].

3.2. Under Variable Precision Rough Sets

In the classical rough set model, L- and U-reducts are essential and we can evaluate attribute importance corresponding to them. In the VPRS model, we have 10 meaningful structure-enhancing reducts. Then we would like to evaluate attribute importance corresponding to the 10 reducts. To this end, we should define a cooperative game corresponding to each reduct. Therefore the problem is how to define such a game.

We define

\[ \partial^L_\beta(x : A) = \{ D_i \mid x \in \overline{A}_\beta(D_i), \ i = 1, \ldots, p \}, \]

\[ \partial^U_\beta(x : A) = \{ D_i \mid x \in \overline{A}_\beta(D_i), \ i = 1, \ldots, p \}. \]

Using those, we further define the following predicates:

\[ P^L_\beta(x : A) = \"\partial^L_\beta(x : A) \supseteq \partial^L_\beta(x : C)\" \]

\[ P^U_\beta(x : A) = \"\partial^U_\beta(x : A) \supseteq \partial^U_\beta(x : C)\" \]

\[ P^\beta_\beta(x : A) = \"(\partial^U_\beta(x : A) \setminus \partial^U_\beta(x : C)) \subseteq (\partial^L_\beta(x : C) \setminus \partial^L_\beta(x : C))\" \]

\[ P^\beta_\beta(x : A) = \"\partial^U_\beta(x : C) = \emptyset or \partial^L_\beta(x : A) \neq \emptyset\". \]

Then we have the following equivalences:

\( (L) \iff \forall x \in U, P^L_\beta(x : A), (U) \iff \forall x \in U, P^U_\beta(x : A), (B) \iff \forall x \in U, P^\beta_\beta(x : A), (UN) \iff \forall x \in U, P^\beta_\beta(x : A) \).

Utilizing those properties, for a given \( \beta \in [0, 0.5) \), we can define a cooperative game corresponding to each structure-enhancing reduct as follows:

\[ v_L(A : \beta) = \frac{\sum_{x \in U} |P^L_\beta(x : A)|}{|U|}, \]

\[ v_U(A : \beta) = \frac{\sum_{x \in U} |P^U_\beta(x : A)|}{|U|} \]

\[ v_{\beta\beta}(A : \beta) = \frac{\sum_{x \in U} |P^\beta_\beta(x : A)|}{|U|}. \]
Table 1: A Decision Table

<table>
<thead>
<tr>
<th>obj.</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>a5</th>
<th>a6</th>
<th>(f₀, f₁, f₂)</th>
</tr>
</thead>
<tbody>
<tr>
<td>p₁</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(5, 0, 0)</td>
</tr>
<tr>
<td>p₂</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(1, 0, 0)</td>
</tr>
<tr>
<td>p₃</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(1, 1, 0)</td>
</tr>
<tr>
<td>p₄</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(0, 1, 1)</td>
</tr>
<tr>
<td>p₅</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>(1, 0, 1)</td>
</tr>
<tr>
<td>p₆</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(1, 1, 1)</td>
</tr>
<tr>
<td>p₇</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>(0, 5, 0)</td>
</tr>
<tr>
<td>p₈</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>(0, 0, 5)</td>
</tr>
</tbody>
</table>

\[ v_{L}^β(A:β) = \frac{\sum_{x∈U} |P_B^\beta(x:A) \land P_D^\beta(x:A)|}{|U|}, \]

\[ v_{L\cup\cup}(A:β) = \frac{\sum_{x∈U} |P_B^\beta(x:A) \land P_D^\beta(x:A)|}{|U|}, \]

\[ v_{L\cup\cup}(A:β) = \frac{\sum_{x∈U} |P_B^\beta(x:A) \land P_D^\beta(x:A)|}{|U|}, \]

\[ v_{L\cup\cup}(A:β) = \frac{\sum_{x∈U} |P_B^\beta(x:A) \land P_D^\beta(x:A)|}{|U|}, \]

\[ v_{L\cup\cup}(A:β) = \frac{\sum_{x∈U} |P_B^\beta(x:A) \land P_D^\beta(x:A)|}{|U|}, \]

where \(|P|\) is the truth value of a statement \(P\), i.e., \(|P| = 1\) if \(P\) is true and \(|P| = 0\) otherwise. The attribute importance can be evaluated by Shapley values of those games.

Note that \(v_L(A) = v_L(A:0) - 1 + \gamma_C(D) = v_{L\cup\cup}(A:0) - 1 + \gamma_C(D)\) and \(v_D(A) = v_{L\cup\cup}(A:0) = v_{L\cup\cup}(A:0) = v_{L\cup\cup}(A:0) = v_{L\cup\cup}(A:0) = v_{L\cup\cup}(A:0)\).

4. An Illustrative Example

Consider a decision table given in Table 1. In Table 1, we have six condition attributes \(C = \{a_1, a_2, a_3, a_4, a_5, a_6\}\) and a single decision attribute \(d\). Each condition attribute \(a_i\) takes an attribute value in \(V_{a_i} = \{0, 1\}\) while decision attribute \(d\) takes an attribute value in \(V_d = \{0, 1, 2\}\). Objects are partitioned into eight patterns \(p_j (j = 1, 2, \ldots, 8)\) by condition attribute values. The row of Table 1 does not correspond to an object but a pattern \(p_j\), \(f_i\) stands for the frequency of objects which has a pattern \(p_j\) of condition attribute values and decision attribute value \(i (i = 0, 1, 2)\).

Let \(β = 0.35\). The 10 kinds of reducts to Table 2 are shown in [6]. Calculating attribute importance as Shapley values based on games defined in Subsection 3.2, we obtain Table 2. From Table 2, we know \(a_5\) is most important in many cases. \(a_4\) is also constantly important. On the contrary, \(a_1\) is not very important. As the requirement for the corresponding reduct becomes strong, the importance of \(a_4\) increases while the importance of \(a_6\) roughly decreases. Because the monotonicity of each game is not guaranteed, the importance of \(a_1\) becomes negative for \(v_{L\cup\cup}^\beta\).

As is demonstrated in this example, the attribute importance can vary with the required level of structure enhancement. If selection of structure enhancement is difficult, the attribute importance should be evaluated from many viewpoints. Moreover, the selection of \(β\) is also an important problem. From the original purpose of the VPRS model, \(β\) should be selected considering the size of lower approximations and the allowable error level.

Acknowledgments

This work has been partially supported by the Grant-in-Aid for Scientific Research (B) No.17310098.

References

Solutions of Cooperative Fuzzy Games Extended from Ordinary Cooperative Games

Tetsuzo Tanino†
†Graduate School of Engineering, Osaka University
2-1 Yamada-oka, Suita, Osaka 565-0871 JAPAN
Email: tanino@eei.eng.osaka-u.ac.jp

Abstract—Cooperative fuzzy games are interesting and useful since they can deal with partial cooperation of players. In particular, special classes of fuzzy games obtained as extensions of ordinary crisp cooperative games have been studied by several authors. They include the multilinear extension and the Lovász extension. In this paper we focus on those classes of cooperative fuzzy games and their solutions, namely allocation schemes of profits among the players. We note that each crisp cooperative game can be represented as a linear combination of unanimity games with the coefficients called the (Harsanyi) dividends. We define solutions in terms of the dividends and investigate relationships between the proposed solutions and the existing solutions. We also study some properties of the solutions.

1. Introduction

A cooperative game (or coalitional game) is usually described by a set of players and a characteristic function. In a transferable utility game only one number is attached to each coalition. In order to deal with practical situations in cooperative game theory, we have to take into account partial cooperation by players. It has been studied through fuzzy coalitions instead of ordinary coalitions which are subsets of players and leads to cooperative fuzzy games [1]. Each fuzzy coalition is identified with a point in the hypercube $[0,1]^n$, while an ordinary coalition is regarded as a vertex of this hypercube, i.e., a point in $\{0,1\}^n$. Another application of fuzzy theory to cooperative games is in studies of cooperative games with fuzzy worths (coalitional values) [6]. In this paper, however, we deal only with cooperative games with fuzzy coalitions as cooperative fuzzy games.

We focus on cooperative fuzzy games which can be obtained by extending ordinary cooperative games. A special interesting class of those games, including the multilinear extension and the Lovász extension, was discussed in our former papers [11]. In this paper we consider solutions of extended cooperative fuzzy games, particularly those based on the dividends of the original crisp game.

2. Cooperative Fuzzy Games

Let $N$ be a finite set of $n$ elements, i.e., $N = \{1,2,\ldots,n\}$. Elements of $N$ are called players. Any subset of $N$ is called a coalition. A (transferable utility) game on $N$ is a set function $v : 2^N \to \mathbb{R}$ with $v(\emptyset) = 0$. The function $v$ is usually called a characteristic function and each value $v(S)$ is called the worth of the coalition $S$. The set of all games on $N$ is denoted by $\Gamma^N$. We use two kinds of symbols for set inclusion: $S \subseteq T$ means that $S$ is a subset of $T$, while $S \subset T$ implies that $S$ is a proper subset of $T$.

The sum $v + w$ of two games $v,w \in \Gamma^N$ and the scalar multiplication of $v \in \Gamma^N$ by a scalar $\alpha \in \mathbb{R}$ are defined by $(v + w)(S) = v(S) + w(S)$ for all $S \subseteq N$, and $(\alpha v)(S) = \alpha v(S)$ for all $S \subseteq N$. respectively. Thus the space $\Gamma^N$ of all games on $N$ is a vector space and its dimension is clearly $2^n - 1$, since each game is specified by the worths $v(S)$ for all $S \subseteq N$ with $S \neq \emptyset$. As a basis in this space we may consider unanimity games $u_T$ defined by

$$u_T(S) = \begin{cases} 1 & \text{if } S \supseteq T, \\ 0 & \text{otherwise,} \end{cases}$$

for any $T \subseteq N$, $T \neq \emptyset$. Then each game $v \in \Gamma$ is a linear combination of unanimity games, $v = \sum_{\emptyset \neq T \subseteq N} d_T(v)u_T$.

The coefficient $d_T(v)$ is given by

d_T(v) = \sum_{S \subseteq T} (-1)^{|T| - |S|} v(S)

and called the (Harsanyi) dividend of $T$ for the game $v$. For convenience’ sake, we may put $d_{\emptyset}(v) = 0$. Then we may write $v = \sum_{T \subseteq N} d_T(v)u_T$. In combinatorics, $d_T(v)$ viewed as a set function on $2^N \setminus \{\emptyset\}$ is called the Möbius transform of $v$. The dividends satisfy the following recursive formula:

$$d_T(v) = \begin{cases} 0, & \text{if } T = \emptyset, \\ v(T) - \sum_{S \subseteq T} d_S(v), & \text{if } T \neq \emptyset. \end{cases}$$
In a cooperative game, the coalition $S$ is identified with the vector $e^S$, defined by $e^S_i = 1$ if $i \in S$ and $e^S_i = 0$ otherwise, and the domain $2^N$ of the characteristic function $v$ is identified with $\{0, 1\}^n$, i.e., $v : \{0, 1\}^n \to \mathbf{R}$. Hence extending $\{0, 1\}^n$ to $[0, 1]^n$ implies extending ordinary coalitions to fuzzy coalitions. Thus, given the player set $N$, a cooperative fuzzy game $\xi$ on $N$ is a function from $[0, 1]^n$ to $\mathbf{R}$ with $\xi(0) = 0$. The set of all cooperative fuzzy games on $N$ is denoted by $\Delta^N$.

In this paper, for $s, t \in [0, 1]^n$, vectors $s \lor t$ and $s \land t \in [0, 1]^n$ are defined by

$$(s \lor t)_i = \max\{s_i, t_i\}, \quad (s \land t)_i = \min\{s_i, t_i\}, \quad i = 1, \ldots, n.$$ 

For $s \in [0, 1]^n$, let supp $s = \{i \in N \mid s_i > 0\}$.

**Definition 1** A cooperative fuzzy game $\xi \in \Delta^N$ is said to be

1. **weakly superadditive** if $\xi(s) + \xi(t) \leq \xi(s \lor t)$ for all $s, t \in [0, 1]^n$ such that $s \land t = 0$.
2. **strongly superadditive** if $\xi(s) + \xi(t) \leq \xi(s + t)$ for all $s, t \in [0, 1]^n$ such that $s + t \in [0, 1]^n$.
3. **convex** if $\xi(s) + \xi(t) \leq \xi(s \lor t) + \xi(s \land t)$ for all $s, t \in [0, 1]^n$.

It is obvious that if the game $\xi$ is strongly superadditive, then it is weakly superadditive. If $\xi$ is convex, it is weakly superadditive. In Brânzei et al. [3], a fuzzy game is said to be convex if it satisfies the coordinate-wise convexity condition in addition to the inequality in the above definition.

The sum of two games $\xi, \eta \in \Delta^N$, and the scalar multiplication of $\xi$ by $\alpha \in \mathbf{R}$ is defined by

$$(\xi + \eta)(s) = \xi(s) + \eta(s), \quad \forall s \in [0, 1]^n,$$

$$(\alpha \xi)(s) = \alpha \xi(s), \quad \forall s \in [0, 1]^n,$$

respectively. Thus $\Delta^N$ is also a linear space.

**3. Cooperative Fuzzy Games as Extensions of Ordinary Cooperative Games**

A typical cooperative fuzzy game is obtained from an ordinary cooperative game $v \in \Gamma^N$ by extending $v$ in an appropriate manner. In the following we denote by $\xi_v \in \Delta^N$ a cooperative fuzzy game obtained by extending $v \in \Gamma^N$. Since $\xi_v$ is an extension of $v$,

$$\xi_v(e^S) = v(S), \quad \forall S \subseteq N.$$ 

When we assume that the way of extension, i.e., the mapping which associates $\xi_v \in \Delta^N$ with each $v \in \Gamma^N$ is linear, we say that the obtained extension $\xi_v$ is a U-extension of $v$. If $\xi_v$ is a U-extension, it can be represented as a linear combination of the U-extensions $\xi_{wT}$ of the unanimity games $w_T$ as

$$\xi_v = \sum_{T \subseteq N} d_T(v)\xi_{wT}.$$ 

Moreover, we assume two additional conditions, restriction invariance and monotonicity, on extensions to obtain W-extensions. The exact definitions of U-extensions and W-extensions are given as follows:

**Definition 2** The extension $\xi_v$ of $v$ is said to be a U-extension of $v$ if the way of extension is linear, i.e.,

$$\xi_{v_1 + v_2} = \xi_{v_1} + \xi_{v_2}, \quad \forall v_1, v_2 \in \Gamma^N,$$

$$\xi_{\alpha v} = \alpha \xi_v, \quad \forall v \in \Gamma^N, \forall \alpha \in \mathbf{R}.$$ 

A U-extension is said to be a W-extension if for any nonempty $T \subseteq N$,

$$\xi_{w_T}(s) = \xi_{w_T}(s_{\mid T}), \quad \forall s \in [0, 1]^n,$$

$$\xi_{w_T}(s) \leq \xi_{w_T}(t), \quad \forall s, t \in [0, 1]^n$$ 

such that $s \leq t$.

Here $s_{\mid T} \in [0, 1]^n$ is defined by $(s_{\mid T})_i = s_i$ if $i \in T$ and $(s_{\mid T})_i = 0$ otherwise.

**Remark 1** Let $\xi_{w_T}$ be a W-extension of $w_T$. Note that $0 \leq s \leq \text{supp } s$ for any $s \in [0, 1]^n$. Therefore, if $T \not\subseteq \text{supp } s$, then

$$0 = \xi_{w_T}(0) \leq \xi_{w_T}(s) \leq \xi_{w_T}(\text{supp } s) = w_T(\text{supp } s) = 0.$$ 

Hence $\xi_{w_T}(s) = 0$ for $T \not\subseteq \text{supp } s$.

Thus, we obtain a stronger result for linear representation of a W-extension.

**Proposition 1** [8] If $\xi_v \in \Delta^N$ is a W-extension of $v \in \Gamma^N$, then

$$\xi_v(s) = \sum_{T \subseteq \text{supp } s} d_T(v)\xi_{w_T}(s), \quad \forall s \in [0, 1]^n.$$ 

Two well-known examples of W-extensions are the multilinear extension $m_v$ introduced by Owen [9, 10] and the Lovász extension $l_v$ [5]. They are given by

$$m_{u_T}(s) = \prod_{i \in T} s_i \quad \text{and} \quad l_{w_T}(s) = \min_{i \in T} s_i.$$ 

Some properties of these extensions such as superadditivity and convexity are given in Bilbao [2] and Tanino [11].
4. Solutions for U-Extensions of Ordinary Cooperative Games

In this section we consider solutions for cooperative fuzzy games obtained as U-extensions from ordinary cooperative games. A point-valued solution of fuzzy games is a function \( f : [0, 1]^n \times \Delta^N \to \mathbb{R}^n \) which associates an \( n \) dimensional vector with each pair of a fuzzy coalition \( s \) and a fuzzy game \( \xi \). We assume linearity of \( f \) with respect to \( \xi \). Then for each \( v \in \Gamma^N \), \( s \in [0, 1]^n \) and \( i \in N \),

\[
  f_i(s, \xi_v) = f_i(s, \xi) \sum_{T \subseteq N} d_T(v) u_T \]

\[
  = f_i(s, \sum_{T \subseteq N} d_T(v) \xi_{u_T})
  = \sum_{T \subseteq N} d_T(v) f_i(s, \xi_{u_T}).
\]

Therefore, this value is specified by values of \( f_i(s, \xi_{u_T}) \).

When a fuzzy coalition \( s \) is formed, the players participating in this coalition will share the worth \( \xi_{u_T}(s) \) for the extension of the unanimity game \( \xi_{u_T} \). We introduce a sharing system \( p = (p^T_i)_{T \subseteq N, i \in T} \) satisfying \( p^T_i \neq 0 \) and \( \sum_{i \in T} p^T_i = 1 \). We put

\[
  f_i(s, \xi_{u_T}) = \begin{cases} 
    p^T_i \xi_{u_T}(s), & \text{if } i \in T, \\
    0, & \text{if } i \notin T.
  \end{cases}
\]

Hence the value corresponding to the sharing system \( p \) is given by

\[
  f_i(s, \xi_{u_T}) = \sum_{T \subseteq N, T \ni i} d_T(v) p^T_i \xi_{u_T}(s) \quad \text{for each } i \in N.
\]

In particular, the case \( p^T_i = \frac{1}{|T|} \) for all \( i \in T \) corresponds to the Shapley value and the obtained value is

\[
  \phi_i(s, \xi_v) = \sum_{T \subseteq N, T \ni i} \frac{d_T(v)}{|T|} \xi_{u_T}(s) \quad \text{for each } i \in N.
\]

When \( \xi_v = l_v \) (Lovász extension),

\[
  \phi_i(s, \xi_v) = \sum_{T \subseteq N, T \ni i} \frac{d_T(v)}{|T|} \min_j s_j \quad \text{for each } i \in N.
\]

It is clear that all these values are zero for \( i \notin \text{supp } s \).

The above Shapley value for the Lovász extension was studied by Tsurumi et al. [13] and also Moritani et al. [7]. In Tsurumi et al. [13], the explicit formula is slightly different, they proposed the following Shapley value for \( v \in \Gamma^N \) and \( s \in [0, 1]^n \):

\[
  g_i(s, l_v) = \int_0^1 \varphi_i([s]_h, v) dh, \quad i \in N,
\]

where \( [s]_h = \{ j \in N \mid s_j \geq h \} \), \( h \in [0, 1] \) and \( \varphi_i([s]_h, v) \) is the Shapley value of \( i \in N \) for the subgame \( v||[s]_h \) of \( v \) on \( [s]_h \), i.e., in terms of dividends

\[
  \varphi_i([s]_h, v) = \sum_{T \subseteq [s]_h, T \ni i} \frac{d_T(v)}{|T|}.
\]

We should note that \( T \subseteq [s]_h \) if and only if \( s_j \geq h \) for all \( j \in N \), i.e., if and only if \( \min_{j \in T} s_j \geq h \). Therefore we can prove that \( g_i(s, l_v) = \phi_i(s, l_v) \) for any \( i \in N \).

Moritani et al. [7] defined the induced Shapley value and proved that it coincides with the Shapley value by Tsurumi et al. [13]. This fact will be discussed more generally in the next section.

The most fundamental set-valued solution for cooperative games is the core. Tsurumi et al. [12] defined the core of a cooperative fuzzy game \( \xi \) depending on the fuzzy coalition \( s \):

\[
  C(s, \xi) = \{ x \in \mathbb{R}^n \mid \sum_{i \in \text{supp } s} x_i = \xi(s), x_i = 0, \forall i \notin \text{supp } s, \sum_{i \in \text{supp } t} x_i \geq \xi(t), 0 \leq t \leq s \}.
\]

5. Relationships with Solutions of Induced Games

In this section we consider induced games defined from a cooperative fuzzy game, which are slightly different from but essentially the same as those in [7].

**Definition 3** Let \( \xi \in \Delta^N \) be a cooperative fuzzy game and \( s \in [0, 1]^n \) be a fuzzy coalition. Then a crisp game \( w^s_\xi \in \Gamma^N \) defined by

\[
  w^s_\xi(T) = \xi(s|T), \quad \text{for all } T \subseteq N
\]

is called the induced game from \( \xi \) and \( s \).

As for the convexity of the induced game we can prove the following proposition easily as in [7].

**Proposition 2** If a cooperative fuzzy game \( \xi \in \Delta^N \) is convex, then for any fuzzy coalition \( s \in [0, 1]^n \) the induced game \( w^s_\xi \in \Gamma^N \) from \( \xi \) and \( s \) is convex as a crisp cooperative game.

If \( \xi_v \) is a U-extension of \( v \in \Gamma^N \), then the induced game from \( \xi_v \) and \( s \) is

\[
  w^s_{\xi_v}(T) = \xi_v(s|T) = \sum_{R \subseteq N} d_R(v) \xi_{u_R}(s|T), \quad T \subseteq N.
\]

**Proposition 3** Let \( v \in \Gamma^N \), \( \xi_v \in \Delta^N \) be a W-extension of \( v \) and \( s \in [0, 1]^n \) be a fuzzy coalition. Then, for all nonempty \( T \subseteq N \),

\[
  d_T(w^s_{\xi_v}) = d_T(v) \xi_{u_T}(s).
\]
Remark 2 In view of Remark 1, $d_T(w^s_{\xi}) = 0$ for any $T \subseteq \text{supp } s$.

Moritani et al. [7] defined the induced Shapley value for a cooperative game $\xi \in \Delta^N$ and a fuzzy coalition $s \in [0,1]^n$, as the Shapley value of the induced game $w^s_{\xi}$. Analogously, we define

$$\Phi_i(s,\xi) = \sum_{T \subseteq N, T \ni i} \frac{d_T(w^s_{\xi})}{|T|}, \quad i \in N.$$  

More generally, for the sharing system $p = (p^T_i)_{T \subseteq N, i \in T}$, we may define the payoff vector

$$F^p_i(s,\xi) = \sum_{T \subseteq N, T \ni i} p^T_i d_T(w^s_{\xi}), \quad i \in N.$$  

In view of Proposition 5, we directly obtain the following theorem, which was proved only for the Shapley value of the Lovász extension in [7].

Theorem 1 Let $v \in \Gamma^N$, $\xi_v \in \Delta^N$ be a W-extension of $v$ and $s \in [0,1]^n$ be a fuzzy coalition. Then, for any sharing system $p = (p^T_i)_{T \subseteq N, i \in T}$,

$$f^p_i(s,\xi_v) = F^p_i(s,\xi_v), \quad \forall i \in N.$$  

In particular,

$$\phi_i(s,\xi_v) = \Phi_i(s,\xi_v), \quad \forall i \in N.$$  

6. Some properties of solutions

In this section we discuss some properties of the solution $f^p(s,\xi_v)$.

Proposition 4 Let $v \in \Gamma^N$, $\xi_v \in \Delta^N$ be a U-extension of $v$ and $s \in [0,1]^n$ be a fuzzy coalition. Then, for any sharing system $p = (p^T_i)_{T \subseteq N, i \in T}$,

$$\sum_{i \in N} f^p_i(s,\xi_v) = \xi_v(s).$$  

Proposition 5 Let $v \in \Gamma^N$, $\xi_v \in \Delta^N$ be a W-extension of $v$ and $s \in [0,1]^n$ be a fuzzy coalition. Then, for any sharing system $p = (p^T_i)_{T \subseteq N, i \in T}$,

$$f^p_T(s,\xi_v) = \left\{ \begin{array}{ll} \sum_{T \subseteq \text{supp } s, T \ni i} d_T(v)p^T_i \xi_v(t) & \text{ if } i \in \text{supp } s, \\ 0 & \text{ if } i \notin \text{supp } s. \end{array} \right.$$  

Proposition 6 Let $v \in \Gamma^N$ be a positive cooperative game, i.e., $d_T(v) \geq 0$ for all $T \subseteq N$, $\xi_v \in \Delta^N$ be a W-extension of $v$ and $s,t \in [0,1]^n$ be fuzzy coalitions such that $s \leq t$. Then, for any sharing system $p = (p^T_i)_{T \subseteq N, i \in T}$,

$$f^p_i(s,\xi_v) \leq f^p_i(t,\xi_v).$$  

Theorem 2 Let $v \in \Gamma^N$ be a positive cooperative game, $\xi_v \in \Delta^N$ be a W-extension of $v$, $s \in [0,1]^n$ be a fuzzy coalition and $p$ be a sharing system. Then

$$f^p(s,\xi_v) \in C(s,\xi_v).$$  

Acknowledgments

This research is supported by the Japan Society for the Promotion of Science under the Grant-in-Aid for Scientific Research No. 16510114.

References

A Solution for Fuzzy Generalized Multialternative Games

Masayo Tsurumi†, Masahiro Inuiguchi‡ and Tetsuzo Tanino‡
†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
‡Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871 JAPAN
Email: {tsurumi, inuiguti}@sys.es.osaka-u.ac.jp, tanino@eee.eng.osaka-u.ac.jp

Abstract—In this paper, we shall propose a fuzzy generalized multialternative games. This game is obtained through the following two steps. First, we define a generalized multialternative game which enables us to discuss situations where players have multiple alternatives and each player chooses an alternative or nothing. Second, as a fuzzy extension of generalized multialternative games, a fuzzy generalized multialternative game (an FGM game) is proposed. By FGM games, we can deal with a partial cooperation to each alternative. A fuzzy marginal contribution for an FGM game is proposed. As a solution for FGM games, we define the expected value of marginal contribution when all partial cooperative relations arise with equal probability. As a special case of FGM games, we define the multilinear extension of a generalized multialternative game. Properties of our solution in this case is discussed.

1. Introduction

There are many situations where multiple decision makers (players) have multiple alternatives. Such situations have been formulated as multialternative games or games with r alternatives [4, 5, 6]. Conventional multialternative games are based on the proceeds when each player chooses one among alternatives. However, in many situations, some players do not choose any alternatives or some players choose several alternatives to certain extents.

In this paper, we shall formulate situations where some players choose one alternative or none as generalized multialternative games. Furthermore, we shall define games for situations where some players choose some alternatives to certain extents as fuzzy generalized multialternative games (FGM games). For FGM games, we shall give a definition of fuzzy-type marginal contribution (fuzzy contribution). The expectation of fuzzy contribution when all fuzzy cooperative relations (fuzzy arrangements) are equally likely to arise is considered. We will call it the Banzhaf-type fuzzy marginal contribution and regard it as a solution for FGM games. Moreover, we introduce the multilinear extension of generalized alternative games. We regard the set of them as a class of FGM games and discuss the Banzhaf-type fuzzy marginal contribution on this class.

2. Conventional cooperative games and their multilinear extension

Let \( N = \{1, 2, \ldots, n\} \) be the set of all players. Then \( v : 2^N \rightarrow \mathbb{R} \) satisfying \( v(\emptyset) = 0 \) is called a conventional cooperative game or simply a game. Equivalently, a conventional cooperative game is represented by \( v : \{0, 1\}^N \rightarrow \mathbb{R} \) satisfying \( v(0, \ldots, 0) = 0 \). The set of all conventional cooperative games is denoted by \( G \).

A function \( g : G \rightarrow \mathbb{R}^N \) can be regarded as a solution for conventional cooperative games. The Banzhaf value is one of major solutions and is defined as follows;

**Definition 1** [3] A function \( \beta : G \rightarrow \mathbb{R}^N \) is called the Banzhaf value if for any \( v \in G \) and \( i \in N \),

\[
\beta_i(v) = \frac{1}{2^{n-1}} \sum_{S \subseteq N \setminus \{i\}} [v(S \cup \{i\}) - v(S)].
\]

A function \( h : \mathbb{R}^n \rightarrow \mathbb{R} \) is called a multilinear function if there exist constants \( C_S (S \subseteq N) \) such that \( h(p_1, \ldots, p_n) = \sum_{S \subseteq N} C_S \prod_{j \in S} p_j \). Let us define \( a_i^S = (a_{i_1}^S, \ldots, a_{i_r}^S) \in \{0, 1\}^r \) by \( a_i^S = 1 \) for any \( i \in S \) and \( a_i^S = 0 \) otherwise. Then the following theorem holds.

**Theorem 1** [7] A multilinear function satisfying \( h(a^T) = v(T) \) for any \( T \subseteq N \) is unique and represented by

\[
h_i(p_1, \ldots, p_n) = \sum_{S \subseteq N} \left[ \prod_{j \in S} p_j \prod_{j \notin S} (1 - p_j) \right] v(S).
\]

The function (1) is called the multilinear extension (MLE) of \( v \). Let \( h_i^*(p_1, \ldots, p_n) \) denote the partial derivative of \( h \) with respect to \( p_i \). Then the following holds;

\[
h_i^*(p_1, \ldots, p_n) = \sum_{S \subseteq N, S \ni \{i\}} \prod_{j \in S} p_j \prod_{j \notin S} p_j [v(S) - v(S \setminus \{i\})].
\]

The following theorem holds.

**Theorem 2** [8] The following holds for any \( v \in G \);

\[
\beta_i(v) = \int_0^1 \cdots \int_0^1 h_i^*(p_1, \ldots, p_n) dp_n \cdots dp_1 = h_i^*(\frac{1}{2}, \ldots, \frac{1}{2}).
\]
For the class of voting games which is a formulation for voting situations, Strafﬁn[9] have given an interpretation of
Theorem 2.
A function \( f : [0, 1]^N \rightarrow \mathbb{R} \) satisfying \( f(0, \ldots, 0) = 0 \) is
called a (conventional) fuzzy game \([1, 2, 10, 11, 12, 13]\).
Note that the domain is \([0, 1]^N\) while the domain of a
conventional cooperative game is \([0, 1]^N\). An element of
\([0, 1]^N\) is called a (conventional) fuzzy coalition.
We shall give another interpretation of Theorem 2 as follows;

Remark 1 By restricting the domain of the MLE \( h^r \) to
\([0, 1]^p\), we obtain a cooperative fuzzy game. Then the partial derivative \( h^r_i \) of \( h^r \) with respect to \( p_i \) can be regarded
as a fuzzy-type of marginal contribution of the player \( i \)
(fuzzy contribution). Moreover, Theorem 2 means that the
Banzhaf value is the expected value of fuzzy contribution when all fuzzy coalitions are equally likely to arise. Fur-
thermore, the Banzhaf value coincides with the fuzzy con-
tribution at the center of gravity of the set of all fuzzy coalitions.

3. Multialternative games

Conventional multialternative games deal with situation
where each player chooses just one alternative. Let \( R = \{1, 2, \ldots, r\} \) be the set of all alternatives. These games are
based on arrangements which are defined as follows;

Definition 2 A partition \( Z = (Z_k)_{k \in R} \) of \( N \) is called an
arrangement, i.e., an arrangement is represented by \( Z = (Z_k)_{k \in R} \) such that
1. \( Z_k \subseteq N, \forall k \in R, \)
2. \( Z_k \cap Z_l = \emptyset, \forall k, l \in R; k \neq l, \)
3. \( \bigcup_{k \in R} Z_k = N. \)

The set of all arrangement is denoted by \( A_r \) or \( A. \)

Based on arrangements, multialternative games are defined as follows;

Definition 3 \([4, 5, 6]\) A function \( m^r : A_r \rightarrow \mathbb{R} \) is called a
multialternative game with \( r \) alternatives if
\[
m^r(Z) = (m^r_1(Z), \ldots, m^r_r(Z)), \]
\[
m^r_i(Z) = 0, \forall k \in R \text{ s.t. } Z_k = \emptyset. \]
The set of all multialternative games with \( r \) alternatives is
denoted by \( MG^r_r \) or \( MG^r. \)

For any multialternative game \( m^r \), \( m^r_i(Z) \) represents the
proceeds of \( Z_k \) in the arrangement \( Z. \)

A \( r \) is isomorphic to \( A^r_0 = \{s = ((s_{i,j})_{r \in R})_{i \in N} \in (0, 1)^{R \times N} | \sum_{j=1}^{r} s_{i,j} = 1, \forall i \in N\}. \)
Hence, a multialternative game is represented by \( m^r : A^r_0 \rightarrow \mathbb{R} \) satisfying \( m^r_i(s) = 0 \) for
any \( k \in R, \sum_{i \in N} s_{i,j} = 0. \) If \( R = \{1\} \), then \( A^r_0 = \{s = (s_{i,j})_{r \in R})_{i \in N} \in (0, 1)^{R \times N} | s_{i,1} = 1, \forall i \in N\} = \{(1, \ldots, 1)\}. \)
This is not isomorphic to the domain of a conventional cooperative
game, i.e., \( 2^N. \) If \( R = \{1, 2\} \), then \( A^r_0 = \{s = (s_{i,j})_{r \in R})_{i \in N} \in (0, 1)^{R \times N} | s_{i,1} = 1, s_{i,2} = 1, \forall i \in N\} = \{(s_{i,1}, s_{i,2}) \in N, \forall i \in N\}. \)
This is isomorphic to the domain of a conventional cooperative
game. For \( R = \{1, 2\} \), a multialternative game is \( m^r : A^r_0 \rightarrow \mathbb{R} \) satisfying \( m^r_i(0, N) = 0 \) and \( m^r_i(N, 0) = 0. \)
Hence, a conventional cooperative game is corresponding to
the first element \( m^r_0 \) of a multialternative game \( m^r \) when
\( |R| = 2. \)

4. Generalized multialternative games

In this paper, we shall formulate generalized multialternative
games as formulations of situations where players choose one alternatives or none. These games are based on
subarrangements which are defined as follows;

Definition 4 \( Y = (Y_k)_{k \in R} \) is called a subarrangement if it satisfies
1. \( Y_k \subseteq N, \forall k \in R, \)
2. \( Y_k \cap Y_l = \emptyset, \forall k, l \in R; k \neq l \)

Denote the set of all subarrangements by \( B_r \) or \( B. \) For \( Y \in
B_r, \) let \( Y_0 = \bigcup_{k \in R} Y_k. \) Let \( B_0 = B \setminus \{(0, \ldots, 0)\} \)

It is also clear that \( A \subseteq B \) for any alternative set \( R. \)
Based on subarrangements, we shall define generalized
multialternative games as follows;

Definition 5 A function \( m^r : B_r \rightarrow \mathbb{R} \) is called a general-
ed multialternative game or a generalized game with \( r \)
alternatives if
\[
m^r(Y) = (m^r_1(Y), \ldots, m^r_r(Y)), \]
\[
m^r_k(Y) = 0, \forall k \in R, \text{ if } \bigcup_{l \in R} Y_l = \emptyset. \]
The set of all generalized games with \( r \) alternatives is de-
noted by \( MG^r_g \) or \( GMG^r. \)

Here, \( m^r_k(Y) \) represents the proceeds of \( Y_k \) for the subarr-
angement \( Y. \)

Note that a generalized multialternative game is obtained by
extending the domain \( A_r \) of a multialternative game to \( B_r \)
and relaxing the second condition.
For any \( Y = (Y_1, \ldots, Y_r) \in B_r, \) we have \( Z = \)
\( (Y_1, \ldots, Y_r, N \setminus \bigcup_{k=1}^r Y_k) \in \cap_{r+1} A_{r+1}. \) For \( m \in GMG^r, \) define
\( m^r : A_{r+1} \rightarrow \mathbb{R}^{+1} \) by \( m^r_i(Z) = m_k(Y) \) for \( k \in R \) and
\( m^r_{r+1}(Z) = 0. \) Then for any \( m \in GMG \) such that \( m_k(Y) = 0 \)
for any \( k \in R, Z_0 = \emptyset, \) the corresponding \( m^r \) can be considered
as a game with \( r + 1 \) alternatives, i.e., \( m^r \in MG^r_{r+1}. \)
\( B_r \) is isomorphic to \( B^r_0 = \{s = ((s_{i,j})_{r \in R})_{i \in N} \in (0, 1)^{R \times N} | \sum_{i=1}^{r} s_{i,j} \leq 1\}. \)
A generalized multialternative game is represented by \( m : B^r_0 \rightarrow \mathbb{R} \) satisfying \( m_k(S) \) for any \( k \in R \) if
\( S_0 = N. \) For \( R = \{1\}, \) a generalized multialternative game coincides with a conventional cooperative game.
Hence, \( GMG^r_1 = \mathbb{G}, \) i.e., conventional cooperative games are special cases of generalized multialternative games.
5. Fuzzy generalized multialternative games (FGM games) and their solution

We shall consider fuzzy generalized multialternative games (FGM games). To introduce them, we shall define fuzzy multialternative subarrangements.

Definition 6 A vector \( s = (s_{i1})_{i \in R}, \ldots, (s_{i_n})_{i \in R} \) \( \in \mathbb{R}^{|N| \times |R|} \) is called a fuzzy subarrangement if
1. \( s_{i_k} \geq 0, \forall k \in R, i \in N \),
2. \( \sum_{k \in R} s_{i_k} \leq 1, \forall i \in N \).

The set of all fuzzy subarrangements is denoted by \( \text{FSA} \).

Note that \( \text{FSA} \) is a standard \( r \)-simplex. For \( r = 1 \), a fuzzy subarrangement is isomorphic to a fuzzy coalition.

Definition 7 A function \( m_f : \text{FSA} \rightarrow \mathbb{R}^+ \) is called a fuzzy generalized multialternative game (an FGM game) if
1. \( m_f(s) = (m_{f1}(s), \ldots, m_{fn}(s)) \) \( \forall s \in \text{FSA} \),
2. \( m_{fk}(s) = 0, \forall k \in R, \) if \( \sum_{i \in N} s_{i_k} = 0 \).

FGM is isomorphic to \( \text{MFA}' = \{ s = ((s_{i1})_{i \in N})_{i \in N} \in (0,1)^{|N|} \mid \sum_{i \in N} s_{i_j} \leq 1, \forall j \in N \} \) while the domain of a generalized multialternative games is isomorphic to \( B' = \{ s = ((s_{i1})_{i \in N})_{i \in N} \in (0,1)^{|N|} \mid \sum_{i \in N} s_{i_j} \leq 1, \forall j \in N \} \). This difference is just the same as the difference between a conventional cooperative game and a fuzzy game. Hence, we consider an FGM game is a natural extension of a generalized multialternative game.

Note that an FGM game coincides with a fuzzy cooperative game when \( r = 1 \).

The set of all FGM games is denoted by \( \mathcal{FGM} \) or \( \mathcal{FGM}_r \). Let us use the notation \( \mathcal{FGM}_r = \{ m_f \in \mathcal{FGM} \mid m_{fk} \in \mathcal{FGM}_r \} \) for any \( k \in R \) and number of discontinuous points of \( \partial m_{fk} \) is finite for any \( i \in N \) and \( k \in R \).

For \( \mathcal{FGM} \subseteq \mathcal{FGM}_1 \), a function \( g : \mathcal{FGM} \to \mathbb{R}^n \) can be considered as a solution on \( \mathcal{FGM} \). We can consider natural properties for a solution \( g : \mathcal{FGM} \to \mathbb{R}^n \).

Property 1 (linearity) For any \( \alpha_1, \alpha_2 \in \mathbb{R} \) and \( m_{f1}, m_{f2}, \alpha_1m_{f1} + \alpha_2m_{f2} \in \mathcal{FGM} \), \( g(\alpha_1m_{f1} + \alpha_2m_{f2}) = \alpha_1g(m_{f1}) + \alpha_2g(m_{f2}) \).

A bijection \( \pi : N \to N \) is called a permutation and \( \pi(i) \in N \) means that the player \( i \) is on the \( \pi(i) \)-th position in the permutation \( \pi \). The set of all permutations is denoted by \( \Pi(N) \). For any \( \pi \in \Pi(N) \), \( s \in \text{MFA} \), \( k \in R \) and \( i \in N \), let \( \pi(s)_{i_k} = s_{\pi(i)_k} \). For any \( \pi \in \Pi(N) \), \( m_f \in \mathcal{FGM} \) and \( k \in R \), define \( \pi m_f \in \mathcal{FGM} \) by \( \pi m_f(s) = m_f(\pi(s)) \) for \( s \in \text{MFA} \).

Property 2 (symmetry) For any \( i \in N, \pi \in \Pi(N), k \in R, m_f \in \mathcal{FGM} \), \( \pi g(m_f) = g(m_f) \).

For \( s \in \text{MFA}, (k, i) \in R \times N \) and \( t \leq 1 - \sum_{j \in R \setminus k} s_{j_k}, \) define \( (s_{i_k}, t) \in \mathcal{MFA} \) by \( (s_{i_k}, t)_{ji} = s_{j_k}, \) for any \( j, i \in R \) satisfying \( j, i \neq (k, i) \) and \( s_{(i_k, t)} = \pi. \) If \( m_f(s_{(i_k, t)}) = 0 \) for any \( 0 \leq t \leq 1 - \sum_{j \in R \setminus k} s_{j_k} \) and \( s \in \text{MFA} \), then the player \( i \) is called a \( k \)-null player.

Property 3 (k-null player property) If a player \( i \) is a \( k \)-null player in \( m_f \in \mathcal{FGM} \), then \( g_{ik}(m_f) = 0 \).

Let us define a new solution based on the Banzhaf value. For \( m_f \in \mathcal{FGM}_1 \), let us use the notation \( m_{f_k} = \partial m_{f_k} / \partial s_{i_k} \) for any \( (k, i) \in R \times N \). Then \( m_{f_k} \) represents the rate of change of the proceeds when the player \( i \) change to choose the alternative \( k \) from choose nothing. We call \( m_{f_k} \) fuzzy-type marginal contribution (fuzzy contribution) of the player \( i \) for the alternative \( k \).

Proposition 1 The expectation of fuzzy contribution of the player \( i \) for the alternative \( k \) when each player chooses a fuzzy subarrangement independently with the uniform distribution is \( (r)^n \int_{\text{MFA}} m_{f_k}(s)ds. \)

Then we shall define the following solution.

Definition 8 For \( \mathcal{FGM}' \subseteq \mathcal{FGM}_1 \), a function \( \beta : \mathcal{FGM}' \to \mathbb{R}^n \) is called the Banzhaf-type expectation of fuzzy contribution on \( \mathcal{FGM}' \) if for any \( (k, i) \in N \times R \) and for any \( m_f \in \mathcal{FGM}', \beta_{ik}(m_f) = (r)^n \int_{\text{MFA}} m_{f_k}(s)ds. \)

It is clear that the Banzhaf-type expectation of fuzzy contribution on \( \mathcal{FGM}' \) is a solution on \( \mathcal{FGM}' \).

6. Multilinear extension of generalized multialternative games

To introduce the multilinear extensions of a generalized multialternative game, let us give a multilinear function.

Definition 9 A function \( f : \mathbb{R}^n \to \mathbb{R} \) is called a multilinear function if there exists constants \( C_z = (S_{i_1})_{i_1 \in R} \in \mathbb{B}_0 \) such that
\[
\phi = \sum_{S_{i_1} \in \mathbb{B}_0} C_z \prod_{i_1 \in R} p_{i_1}. \]

For \( i \in N, k \in R, \pi = (S_{i_1})_{i_1 \in R} \in \mathbb{B}_0 \), define \( p_{i_1}^{\pi} = 1 \) if \( i_1 \in S_k \) and \( p_{i_1}^{\pi} = 0 \) otherwise.

Theorem 3 For \( m \in \mathcal{GM} \) and \( k \in R \), a multilinear function \( f^m : \mathbb{R}^n \to \mathbb{R} \) such that \( f^m((p_{i_1}^{\pi})_{i_1 \in R}) = m(S) \) for any \( S = (S_{i_1})_{i_1 \in R} \in \mathbb{B}_0 \) is unique and given as follows:
\[
f^m = \sum_{S_{i_1} \in \mathbb{B}_0} \prod_{i_1 \in R} p_{i_1}^{\pi} \prod_{j = 1}^n (1 - \sum_{i_1 = 1}^n p_{i_1}^{\pi})^{m_k(S)}. \]

We call the function (2) the multilinear extension of \( m \in \mathcal{GM} \) for the alternative \( k \). We can define a FGM game \( f^m : \mathcal{FMA} \to \mathbb{R}^n \) by \( (f^m)_i(s) = f^m_i(s) \) for any \( k \in R \) and \( s \in \text{MFA} \). Conversely, restricting the domain of the obtained \( f^m \) to \( \mathbb{B}_0 \), then we have \( m \in \mathcal{GM} \). It is noted that there is a one-to-one correspondence between a generalized multialternative game and an FGM game. Let us denote
by $\mathcal{F}G\mathcal{M}_{MLE}$ the class of all such FGM games. Note that $\mathcal{F}G\mathcal{M}_{MLE} \subseteq \mathcal{F}G\mathcal{M}_1$.

Define $\{i\}_k = ((\{i\}_k)_l)_l \in B_0$ by $(\{i\}_k)_l = \{i\}$ for $l \leq k$ and $(\{i\}_k)_l = \emptyset$ otherwise.

**Proposition 2** For any $m_f \in \mathcal{F}G\mathcal{M}_{MLE}$, the Banzhaf-type expectation of fuzzy contribution is represented by

$$\beta_{i,k}(m_f) = \frac{1}{(r+1)^{k-1}} \sum_{S \in \mathcal{R}_k \cap \mathcal{S} \cap \mathcal{N}} \{m(S \cup \{i\}_k) - m(S)\},$$

where $m \in \mathcal{G}\mathcal{M}_\mathcal{R}$ is the generalized multialternative game corresponding to $m_f \in \mathcal{F}G\mathcal{M}_{MLE}$.

It is noted that the Banzhaf-type expectation of fuzzy contribution coincides with the Banzhaf value for a conventional game when $r = 1$.

The following hold for the Banzhaf-type expectation of fuzzy contribution for $m_f \in \mathcal{F}G\mathcal{M}_{MLE}$.

**Theorem 4** For any $m_f \in \mathcal{F}G\mathcal{M}_{MLE}$, we have

$$\beta_{i,k}(m_f) = m_f(S) \left(\left(1 \right)_{j=1}^{r+1} \cdots \left(1 \right)_{k=1}^{r+1}\right).$$

Namely, for any $m_f \in \mathcal{F}G\mathcal{M}_{MLE}$, the Banzhaf-type expectation of fuzzy contribution coincides with the fuzzy contribution at the center of gravity of MFA.

Furthermore, we have the following property of the Banzhaf-type expectation of fuzzy contribution.

**Theorem 5** The Banzhaf-type expectation of fuzzy contribution on $\mathcal{F}G\mathcal{M}_{MLE}$ satisfies linearity, symmetry, $k$-null player property.

7. Conclusion

In this paper, we have formulated situations where some players do not choose any alternatives as generalized multialternative games. As games for situations where some players do not choose any alternatives, fuzzy generalized multialternative games have been defined. For fuzzy generalized multialternative games, we have given a definition of fuzzy-type marginal contribution. Based on it, we have defined the expectation of fuzzy-type marginal contribution when all fuzzy arrangements are equally likely to arise as the Banzhaf-type fuzzy marginal contribution. It can be regarded as a solution for fuzzy generalized multialternative games. Moreover, we have introduced the multilinear extension of generalized alternative games. The set of them can be considered as a class of fuzzy multialternative games. Discussions of the Banzhaf-type fuzzy marginal contribution on this class are given.

As a future research, we shall define the Shapley-type fuzzy marginal contribution. Furthermore, we shall give an axiomatization of the Banzhaf-type and the Shapley-type fuzzy marginal contribution will be discussed.

Acknowledgments

This research is partially supported by Grant-in-Aid for Young Scientists (B), The Ministry of Education, Culture, Sports, Science and Technology, Japan, No. 16710113.

References


Property of Random Series from Chaos Neural Network

Takahiro Nakanishi†, Hitoaki Yoshida† and Satoshi Kawamura‡

†Super Computing and Information Sciences Center, Iwate University
18-8 Ueda 3, Morioka, Iwate, 020-8553 JAPAN
‡Faculty of Science and Engineering, Ishinomaki Senshu University
1 Shinmito, Minamisakai, Ishinomaki, Miyagi 986-8580, JAPAN
Email {takahiro, hitoaki}@iwate-u.ac.jp, kawamura@isenshu-u.ac.jp

Abstract—We have presented a pseudo random number generator using a chaos neural network (CNN) that consists of conventional neurons, and also have presented application of this pseudo random number generator to stream cipher.

In this paper, we report a relation between parameters of CNN and a property of random series from CNN, and also report a relation between a difference in a way to extract of random number from CNN output and property of random number.

1. Introduction

We have continuously studied on the chaos neural network (CNN) that consists of conventional neurons (CNN) [1]-[4]. We have presented a pseudo random number generator using a chaos neural network that consists of conventional neurons (CNP), and also have presented application of this pseudo random number generator (CNP) to stream cipher (CNPC)[5].

Pseudo random numbers generated by CNP are suitable for stream cipher, since they have uniformity and independence. In order to generate the random number of good character by CNP, coefficients of neural network need a tuning.

In this paper, we report a relation between parameters of CNN and a property of random series from CNN, and also report a relation between a difference in a way to extract of random number from CNN output and property of random number.

2. Chaos Neural Network

A chaos neural network that composed of 4 neurons is used for a chaos generator (Figure 1). The neuron is conventional artificial neuron (Figure 2).

The internal state of \( j \)th neuron at time \( t \) is defined as

\[
  u_j(t) = \sum_{i=1}^{n} w_{ij} x_j(t-1) - \theta_j + I_j
\]

(1)

where \( w_{ij} \) is a synaptic weight between \( i \)th neuron and \( j \)th neuron, \( x_j \) is an input from \( j \)th neuron, \( \theta_j \) is a threshold of \( j \)th neuron, \( I_j \) is an external input of \( j \)th neuron. An output from \( j \)th neuron is defined as follows,

\[
  y_j(t) = \frac{1}{1 + \exp(-u_j(t))}
\]

(2)

Poincare sections across an embedded manifold of C-4nn in time-delay coordinate system is shown in Figure 3. Attractor is reconstructed with same embedding dimension \( m = 4 \) and embedding lag \( \tau = 1 \). A maximum Liapunov exponent is 1.7 and correlation dimension is 0.225.

Figure 1: CNN having cyclic structure (C-4nn)
Figure 2: Neuron Model
Figure 3: Poincare section of \( N_1 \)'s output (\( I_1 = I_2 = 0.497 \))
3. Extraction of Pseudo Random Number from Chaos Output

3.1. Extraction method1

We have proposed a method of generating random numbers from CNN outputs which is representing double precision floating-point according to the IEEE 754 floating-point standard. In this method, Lower $L$ bits in fraction bit field(52bit) is extracted from a CNN output as random number(Figure 4,5). In this paper, this extraction method is called method1.

3.2. Extraction method2

In method2, random numbers are extracted from CNN output, too. Numeral system of CNN output is different from one of the method1.

In this method, a CNN output is treated as fixed-point number which is created from exponent field and fraction bit field of a neuron output. Random numbers are obtained from lower $L$ bits of fixed-point numbers (Figure 6).

4. Statistical tests of Random Number

Random numbers extracted from outputs of CNN’s every neuron($N_1 \sim N_4$) with various external input($I_1, I_2$) are tested with standard statistical test suites. In this paper, following empirical tests are performed with the DIEHARD test suite [6]. The DIEHARD test is a battery of 18 stringent statistical test. The DIEHARD program written by G.Marsaglia can be found on the Web [7]. In this paper, random numbers are performed with follow 11 tests the following.

- Birthday Spacing Test
- Count-the-1’s Test on a stream of bytes
- The Craps Test: Throw
- Minimum Distance Test
- The Overlapping Sums Test
- The Binary Rank Test for 6x8 Matrices
- The Runs Test: Runs down
- The Runs Test: Runs up
- The Squeeze Test
- The 3DSPheres Test

Results of tests are called “$p$-values” in the DIEHARD test. These values are real number between 0 and 1. In this paper, the rejection area of each test is 5%-level significance of two-sided alternatives. Given that numbers which are test object are random number, rejected rate is closed to 5%.

5. Test Result

5.1. method1

Figure 7,8 show relation between the DELTA and rejected rates with method1 on Count-the-1’s test on stream of bytes (Count1’s test) and the CRAPS test: throw (CRAPS throw test) , where , extracted bit length $L$ is 16, DELTA is gain of the external inputs ($I_1 = I_2$). The external inputs $I_1$ and $I_2$ are defined as Eq.(3) and (4).

\[
I_1 = -0.497 + 0.0001 \text{ DELTA} \\
I_2 = I_1
\]

Figure 9 shows relation between the DELTA and rejected rates with method1 on the CRAPS throw test with extracted bit length $L = 32$.

These figure shows that rejected rates are closed to 5% with almost DELTA. When a rejected rate of one neuron’s output is far from 5% ,the rejected rate of the other neuron’s output is far from 5%, too. Results of other tests (Birthday Spacing Test, Minimum Distance Test, ...) are similar to this result.

According to these results, number sequence from all neurons are random number when number sequence from one neuron is random number with method1.

5.2. method2

Figure 10 shows relation between the DELTA and rejected rates with method2 on Count1’s test. This figure shows that rejected rates of number sequence generated from $N_4$ are 100% with all DELTA , and that rejected rates of $N_4$ are unrelated to rejected rates of the other neuron.

We think that this phenomenon is caused by computational errors in extraction process. Figure 11 shows relation between the DELTA and rejected rates on Count1’s test.
the lowest 1 bit of extracted random number with method2 is discarded avoiding computational errors. In this paper, this extraction method2 with discarding the lowest 1 bit is called method2-2. In this figure, rejected rate of random number which extracted from all neurons outputs are similar.

Figure 12 shows relation between the DELTA and rejected rate with method2-2 on Count1’s test with L = 33. Although the lowest 1 bit is discarded, the test results of N3, N4 differ from the test results of N1, N2.

According to these results, that number sequence generated from a neuron’s outputs is random number does not support that number sequence generated from the other neuron is random number with method2 and method2-2.

6. Conclusion

The relation between parameters of CNN and a property of random series from CNN and the relation between a difference in a way to extract of random number from CNN output and property of random number are reported in this paper.

Empirical test result suggest that the extraction method1, in which neuron outputs are treated as floating point number, is suited for generating random number.

Acknowledgments

I would like to thank Kazuaki Oozeki, Ryuji Gotoh, Kazuo Fujimaki and Masato SAITO for their detailed comments, suggestions and constant support.

Figure 7: Relation between DELTA and rejected rate with extraction method1 (Count-the-1’s Test on stream of bytes, L = 16)

Figure 8: Relation between DELTA and rejected rate with extraction method1 (The Craps Test: Throw, L = 16)

References

Figure 9: Relation between \textit{DELTA} and rejected rate with extraction method1 (The Craps Test: Throw, $L = 32$)

Figure 10: Relation between \textit{DELTA} and rejected rate with extraction method2 (Count-the-1’s Test on stream of bytes, $L = 17$, the lowest 1 bit is discarded.)

Figure 11: Relation between \textit{DELTA} and rejected rate with extraction method2 (Count-the-1’s Test on stream of bytes, $L = 17$, the lowest 1 bit is discarded.)

Figure 12: Relation between \textit{DELTA} and rejected rate with extraction method2 (The Craps Test: Throw, $L = 33$, the lowest 1 bit is discarded.)
Fractal Analysis of Chaos Neural Network Outputs in Transient State and Steady State

Hitoaki YOSHIDA†, Osamu OHIRA‡, Hiroto TAIRA‡, and Takahiro NAKANISHI†

†Super Computing and Information Sciences Center Iwate University
3-18-3 Ueda, Morioka, Iwate, 020-8550, Japan
‡ Graduate School of Computer and Information Sciences, Iwate University
4-3-5 Ueda, Morioka, Iwate, 020-8551, Japan
E-mail: †{takahiro, hitoaki}@iwate-u.ac.jp, {h13j14, h13j39}@cis.iwate-u.ac.jp‡

Abstract—We have continuously studied on the chaos neural networks (CNN) that consist of conventional artificial neurons. In this study, we report fractal analysis of chaos neural network outputs both in a transient state and a steady state. Fractal analysis of the transient state has revealed topological transformation and the route to the steady state (or the strange attractor). Multifractal analysis of the steady state has revealed methods to synthesize new chaos series, computationally.

1. Introduction

We have continuously studied on the chaos neural networks (CNN) that consist of conventional artificial neurons and generate chaotic outputs [1-4]. Recently, we also have applied a CNN to a stream cipher [5-6], and have commercialized a CNN cipher as a security system.

In this paper, we report fractal analysis of the chaos neural network output in both a transient and a steady state. Fractal analysis is expected to afford a key property of the CNN. The ultimate goal of our study is implementation of a universal chaos generator.

2. Chaos Neural Network

A chaos neural network that composed of 4 neurons in discrete-time system is used for a chaos generator (Figure 1). Inputs of jth neuron at time t is defined as

\[ u_j(t) = \sum_{i=1}^{n} w_{ij} x_i(t-1) - \theta_j + I_j \] (1)

where \( w_{ij} \) is a synaptic weight between ith neuron and jth neuron, \( x_i \) is an input from jth neuron, \( \theta_j \) is a threshold of jth neuron, \( I_j \) is an external input of jth neuron. An output from jth neuron (an internal state of jth neuron) is defined as follows:

\[ x_j(t) = \frac{1}{1 + \exp(-u_j(t))} \] (2)

In this study, however, Eq. (3) is used instead of Eq. (1) and Eq. (2) to avoid a computing error originating from cancellation of significant digits.

\[ x_j(t) = \frac{1}{1 + \exp(-u_j(t))} \prod_{i=1}^{4} \exp(-w_{ij} x_i(t-1)) \] (3)

3. Fractal Analysis of Steady State of B-4nn (Part I)

Four internal states of the neurons in B-4nn represent a point, \((x_1(t), x_2(t), x_3(t), x_4(t))\) in four-dimensional (4-D) phase space. The time series of the internal states draws a chaos orbit (Figure 2). The Box-counting method [8-9] gives a fractal dimension of a strange attractor of the chaos in the steady state (after \( t=1000 \) unit time, i.e., \( D(0) = 1.728 \) by using 10000 points \( t=1001-11000 \).

Moreover, Lyapunov spectrum of the strange attractor has also been obtained as \( (0.135, -0.103, -0.403, -0.814) \). The time series is a dissipative system, because Lie derivative [12] has been negative value, -7.37 (an averaged value during 3000 unit time).

4. Fractal Analysis of Transient State of B-4nn

A transient state of B-4nn leading to the strange attractor \( (t = 1-100) \) is analyzed by using Box-counting method [7-8]. Initial values of B-4nn are prepared as
10000 points that are arranged on lattice points of 4-D hyper cube. The set whose members are the 10000 points is called a cloud in the following, and the side length of the 4-D hyper cube corresponds to the cloud size (s). The cloud has transformed nearly same attractor as shown in Figure 2. Box-counting dimension \( D(0) \) of the cloud (10000 points) is investigated with changing the cloud size from \( s = 10^{-5} \) to \( 10^{-8} \) at each unit time.

\[ t_s = -9.3 \log s + 77.0 \]  
(4)

where \( \log \) is a common logarithm.

In the last stage, the curve shaped cloud is folding and tangling, and gradually approaches the strange attractor. It is consistent with well-known Horseshoe map, which is characteristic in a dissipative system. Interestingly, rotational motion of the cloud has been observed during whole transient state. The rotational motion may originate from bias of initial values, the detail, however, will be discussed in the further study.

In the graph of the data without normalization, up-and-down feature of the curve is corresponding to folding and tangling motion of cloud in the 4-D phase space. But as a whole, \( D(0) \) increase gradually. It is corresponding to diffusion of the cloud to entire domain. After the cloud size reaches to the dynamic range, the two curves coincide with each other.

The comparative method between normalized data and data without normalization is useful for distinction the steady state from the transient state, and for investigation of a route to a strange attractor.

5. Border of Transient State and Steady State

The border of the transient state and the steady state is estimated to be \( t_s = 136 \) on the basis of \( D(0) \) value when \( s = 10^{-5} \) (Eq.(4)). But if we assume that the border is the stop time of the rotational motion, the border is estimated to be about \( t = 200 \). The detail will be discussed in the further study. In this study, the values at \( t > 1000 \) is used as the values at a steady state for safety.

6. Fractal Analysis of Steady State of B-4nn (Part II)

6.1. Deterministic Property of Time Series

We have been reported that nature of upper bits and lower bits of the chaotic output from CNN as a fix-point representation system (Figure 4) [3, 4].

Figure 4: Chaotic output is divided into higher bits and lower bits.
After transformation of variables from floating-point representation to fix-point representation, the original chaos time series, \( \{ x_t \}_{t=1}^{T} \), is divided into two time series \( \{ x_t^{up} \} \) and \( \{ x_t^{low} \} \). The former time series consists of upper \( n \) bits of the original time series \( \{ x_t \} \) and the latter consists of lower \( m \) bits of \( \{ x_t \} \). Here,

\[
x_t = x_t^{up} + x_t^{low}
\]

(5)

Otherwise, we describe as follows:

\[
x_t = x_t^{up}(n) + x_t^{low}(m)
\]

(6)

In this work the lowest 1 bit is discarded avoiding computational errors. The time series, \( \{ x_t \}, \{ x_t^{up} \} \) and \( \{ x_t^{low} \} \) are separately analyzed as independent time-series.

In the previous work, randomness of \( \{ x_t^{low} \} \) become clear, and the time series is applied to a pseudo-random-number generator, and to a stream cipher. The result of the recurrence plot analysis [7] suggests that \( \{ x_t^{up} \} \) has deterministic character [3]. The result is shown in Figure 5, by using more precise method than before: averaged values of repeated experiments with different initial conditions. Figure 5 shows relation between the number of the extracted bits and a rate of deterministic points \( (R) \), which is defined as following equation:

\[
R = \frac{N_{RP ∩ IDRP}}{N_{RP}}
\]

(7)

where \( N_{RP} \) is the number of the points in the recurrence plot \( (RP) \), and \( N_{RP ∩ IDRP} \) is the number of the points in \( RP ∩ IDRP \) that is the intersection of RP and the iso-directional recurrence plot \( (IDRP) \) [11]. The IDRP visualizes points with the similar move direction. In previous work, we used a point density, instead of \( R \) [3]. The rate of deterministic points \( (R) \) represents deterministic property of a time series.

Yet, \( \{ x_t^{up} \} \) is still unidentified so far. In this work, \( \{ x_t^{up} \} \) is studied with multifractal analysis. Identification of \( \{ x_t^{up} \} \) will lead implementation of a useful chaos generator.

### 6.2. Multifractal Analysis of \( \{ x_t^{up} \} \)

When a fractal figure is covered with cells (the size of the cell is \( \ell \)), if we express the measure at \( \ell \)th cell as \( P_i(\ell) \), the generalized correlation integral \( (Z_i(q)) \) is written as,

\[
Z_i(q) = \sum_{\ell} [p_\ell(i)]^q
\]

(8)

where \( N(\ell) \) is the number of occupied cells, and \( q \) is a real number. Then the multifractal dimension \( (D(q)) \) is defined as [8, 10]

\[
D(q) = \frac{1}{q-1} \lim_{\ell \to 0} \frac{\log Z_i(q)}{\log \ell}
\]

(9)

In our study, obtained multifractal dimensions are shown in Figure 6. The solid and bold curve indicates \( D(q) \) of the original chaos time series \( \{ x_t \} \), and the solid and thin curve indicates \( D(q) \) of \( \{ x_t^{up}(8) \} \). Both curves show nearly same tendency. However, \( D(q) \) of \( \{ x_t^{low}(24) \} \) changes gently (the dotted curve), probably because of uniformity.

Next, relation between \( D(0), \ D(1), \ D(2), \) and the number of extracted upper bits is studied to determine the boundary condition of the fractal property (Figure 7). All \( D(q) \) values gradually approach the \( D(q) \) value of \( \{ x_t \} \). The border of the fractal property has been 7-8 bits, which is in good agreement with the result in Figure 6. Moreover, the border of deterministic property has been 7-8 bits in Figure 5. Therefore, we concluded that \( \{ x_t^{up}(8) \} \) is a deterministic fractal time series.

![Figure 5: Deterministic Property of \( \{ x_t^{up}(n) \} \) and \( \{ x_t^{low}(m) \} \).](image)

![Figure 6: Obtained Multifractal Dimension, \( D(q) \).](image)
7. Preliminary Study of Chaos Generator

According to the result, relation among three different systems is written as:

\[ x_t = x_t^{up}(8) + x_t^{low}(45) \]  

(10)

Interestingly, it means that the chaos series is described as the linear combination of deterministic fractal series and random number series. The equation shows new chaos generation methods. One possible method is linear combination of newly synthesized \( \{ x_t^{up}(8) \} \) and random number series \( \{ x_t^{low}(45) \} \). New \( \{ x_t^{new}(8) \} \) may generate as completely different fractal series, otherwise may transform by a mapping \( f \):

\[ f : x_t^{up}(8) \rightarrow x_t^{new}(8) \]  

(11)

8. Conclusion

In fractal analysis of CNN outputs in the transient state, the comparative method is useful for distinction the steady state from the transient state, and for investigation of a route to a strange attractor. Multifractal analysis of CNN outputs in the steady state has revealed methods to synthesize new chaos series, computationally. The ultimate goal of our study is implementation of a universal chaos generator.

References

Solving Ability of Hopfield Neural Network for QAP by Changing Chaotic Behavior of Switching Noise

Yoshifumi Tada, Yoko Uwate and Yoshifumi Nishio

Department of Electrical and Electronic Engineering,
Tokushima University, Japan
2-1 Minami-Josanjima Tokushima 770-8506, Japan
Email: \{y-tada, uwate, nishio\}@ee.tokushima-u.ac.jp

Abstract—In our previous research, we confirmed that the chaotic switching noise generated by the cubic map gained a good performance for solving combinatorial optimization problems when the noise was injected to the Hopfield neural network. However, the reason of the good effect of chaotic switching noise has not been clarified completely. In this study, we investigate the solving ability of Hopfield neural network for QAP when the chaotic behavior of the switching noise is changed.

1. Introduction

Combinatorial optimization problems can be solved with the Hopfield neural network (abbr. NN). If we choose connection weights between neurons appropriately according to given problems, we can obtain a good solution by the energy minimization principle. However, the solutions are often trapped into a local minimum and do not reach the global minimum. In order to avoid this critical problem, several people proposed the method adding some kinds of noise for solving traveling salesman problems (TSP) with the Hopfield NN [1]. Hayakawa and Sawada pointed out the chaos near the three-periodic window of the logistic map gains the best performance [2]. They concluded that the good result might be obtained by a property of the chaos noise; short time correlations of the time-sequence. Hasegawa et al. investigated solving abilities of the Hopfield NN with various surrogate noise, and they concluded that the effects of the chaotic sequence for solving optimization problems can be replaced by stochastic noise with similar autocorrelation [3]. We have also studied the reason of the good performance of the Hopfield NN with chaotic noise. We imitated the intermittency chaos by the Gilbert [4] model with 2 states; a laminar state and a burst state. We concluded that the irregular switching of laminar part and burst part is one of the reasons of the good performance of the chaotic noise [5] [6]. Further, we have investigated a performance of chaotic switching noise generated by the cubic map when the noise is injected to the Hopfield NN for quadratic assignment problem (abbr. QAP). We have confirmed that the chaotic switching noise was effective for solving QAP similar to the intermittency chaos noise near the three-periodic window [9]. However, the reason of the good effect of chaotic switching noise has not been clarified completely.

In this study, we investigate solving ability of Hopfield NN for QAP when the chaotic behavior of the switching noise is changed. By computer simulation, we confirm that the network can find good solutions, even when the chaotic behavior is partly replaced by random time series.

2. Solving QAP with Hopfield NN

Various methods are proposed for solving the QAP which is one of the NP-hard combinatorial optimization problems. The QAP is expressed as follow: given two matrices, distance matrix \(C\) and flow matrix \(D\), and find the permutation \(P\) which corresponds to the minimum value of the objective function \(f(P)\) in Eq. (1).

\[
f(P) = \sum_{i=1}^{N} \sum_{j=1}^{N} C_{ij} D_{p(i)p(j)},
\]

where \(C_{ij}\) and \(D_{ij}\) are the \((i,j)\)-th elements of \(C\) and \(D\), respectively, \(p(i)\) is the \(i\)th element of vector \(P\), and \(N\) is the size of the problem. There are many real applications which are formulated by Eq. (1). One example of QAP is find an arrangement of the factories to make a cost the minimum. The cost is given by the distance between the factories and flow of the products between the factories. Other examples are the placement of logical modules in a IC chip, the distribution of medical services in large hospital.

Because the QAP is very difficult, it is almost impossible to solve the optimum solutions in large problems. The largest problem which is solved by deterministic methods may be only 24 in recent study. Further, computation times is very long to obtain the exact optimum solution. Therefore, it is usual to develop heuristic methods which search near optimal solutions in reasonable time.

For solving \(N\)-element QAP by Hopfield NN, \(N \times N\) neurons are required and the following energy function is defined to fire \((i,j)\)-th neuron at the optimal position:

\[
E = \sum_{i,m=1}^{N} \sum_{j,n=1}^{N} w_{im;jn} x_{im} x_{jn} + \sum_{i,m=1}^{N} \theta_{im} x_{im}.
\]
The neurons are coupled each other with the following weight between \((i, m)\)-th neuron and \((j, n)\)-th neuron and the threshold of the \((i, m)\)-th neuron is described by:

\[
w_{im;jn} = -2 \left\{ A(1-\delta_{mn})\delta_{ij} + B\delta_{mn}(1-\delta_{ij}) + \frac{C_{ij}D_{mn}}{q} \right\}
\]

\[
\theta_{im} = A + B
\]

where \(A\) and \(B\) are positive constant, and \(\delta_{ij}\) is Kroneker’s delta. The state of \(N\times N\) neurons are asynchronously up-date due to following difference equation:

\[
x_{im}(t+1) = f \left( \sum_{j,n=1}^{N} w_{im;jn} x_{im}(t) x_{jn}(t) - \theta_{im} + \beta z_{im}(t) \right)
\]

where \(f\) is sigmoidal function defined as follows:

\[
f(x) = \frac{1}{1 + \exp \left( -\frac{x}{\varepsilon} \right)}.
\]

\(z_{im}\) is additional noise, and \(\beta\) limits the amplitude of the noise.

3. Chaotic Switching Noise

In this section, we describe chaotic switching noise injected to the Hopfield NN.

The following cubic map is used to generate the chaotic switching noise.

\[
\hat{y}_{im}(t+1) = -\hat{y}_{im}(t)(\alpha_{c}\hat{y}_{im}^2(t) + 1 - \alpha_{c})
\]

Figure 1 shows the shape of the cubic map Eq. (7). The one-parameter bifurcation diagram of this map is shown in Fig. 2. The attractor becomes symmetric at around \(\alpha_{c} = 3.600\) via an interior crisis. Because the transition of the solution from the positive/negative part to the other part is seldom just after the crisis, the behavior looks like an irregular switching as shown in Fig. 3. We use these time series after the following normalization.

\[
y_{im}(t+1) = \frac{\hat{y}_{im}(t) - \bar{y}}{\sigma_{y}}
\]

where \(\bar{y}\) and \(\sigma_{y}\) are the average and the standard deviation of \(\hat{y}(t)\), respectively.

In our previous research [9], we have confirmed that this chaotic switching noise was effective for solving QAP.

4. Changing Chaotic Behavior of Switching Noise

We consider that the chaotic switching noise (Figs. 3) contains two kinds of chaotic features; one is the switching timing of the upper part and the lower part, the other is the behavior inside the each part.

In the previous research [9], we have confirmed that regular switching between the two parts makes the performance worse. Hence, the former chaotic feature can be
said to be important for solving QAP. How about the latter chaotic feature? In order to answer this question, we replace the chaotic time series inside the each interval of the chaotic switching noise by other time series.

Figure 4 shows three kinds of time series made by replacing the original chaotic time series by chaos, random, and torus. The switch was to multiply the sequence number either by 1 or by $-1$. Please note that we keep the switching timing between the upper part and the lower part.

4.1. Switching Chaos

The logistic map is used to generate the chaotic time series inside the each interval.

$$\hat{l}_{im}(t + 1) = \alpha l\hat{l}_{im}(t)(1 - \hat{l}_{im}(t)).$$

(9)

Varying parameter $\alpha_l$, Eq. (9) behaves chaotically via a periodic-doubling cascade. In this study, we carry out computer simulations using the fully-developed chaos, which is obtained from Eq. (9) for $\alpha_l = 4.0$.

4.2. Switching Random

Random time series are generated by using random function of C compiler. The range of random noise is set to $0 \sim 1$.

4.3. Switching Torus

The sine circle map is used to generate the torus time series inside the each interval.

$$\hat{x}_{im}(t + 1) = \hat{x}_{im}(t) + \alpha x \sin \{6\pi \hat{x}_{im}(t)\} + D.$$  (10)

The parameters are fixed as $\alpha_x=0.04$ and $D=0.1$.

5. Simulation Results

In this section, the simulation results of Hopfield NN with three kinds of switching noise for 12-elements QAP are summarized in Tables 1 and 2. The problem used here was chosen from QAPLIB, whose name is “Nug12.” The global minimum of this target problem is known as 578. The parameters of Hopfield NN are fixed as $A = 0.9$, $B = 0.9$, $q = 140$, $\varepsilon = 0.02$ and $\beta = 0.55$.

The tables show the average of the obtained solutions in 10 trials with different initial conditions, the minimum solution in 10 trials, and the error between the average and the optimal solution calculated by the following equation.

$$\text{Error} = \frac{\text{Ave} - \text{Opt}}{\text{Opt}} \times 100,$$

(11)

where $\text{Opt}$ denotes the optimal solution of the target problem.

From these tables, we can confirm that the switching chaos and the switching random can gain similar performance to the cubic map (chaotic switching noise). This means that the chaotic feature inside the each interval is not very important for solving QAP.

However, we can also notice that the switching torus can not find a good solution. We can conclude that some kinds of irregularity are important inside the each interval of the switching noise.
6. Conclusions

In this study, we have investigated the solving ability of Hopfield NN when the chaotic behavior of the switching noise was replaced by another chaos, random and torus. By computer simulation, we confirmed that the network can find good solution, even when the chaotic behavior was partly replaced by random time series. We conclude that the reason of the good performance of the chaotic switching noise generated from the cubic map is the chaotic switching between the upper part and the lower part, and that chaotic behavior inside the each interval is not important but some kinds of irregularity are necessary.

References


Table 1: Solving abilities for 12-elements QAP. (αc=3.599)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>switching chaos</th>
<th>switching random</th>
<th>switching torus</th>
<th>Cubic Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave</td>
<td>Min</td>
<td>error</td>
<td>Ave</td>
<td>Min</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>-----------</td>
</tr>
<tr>
<td>1000</td>
<td>624.8</td>
<td>586</td>
<td>8.097</td>
<td>645.4</td>
</tr>
<tr>
<td>2000</td>
<td>620.2</td>
<td>586</td>
<td>7.301</td>
<td>631.2</td>
</tr>
<tr>
<td>3000</td>
<td>619.0</td>
<td>586</td>
<td>7.903</td>
<td>627.6</td>
</tr>
<tr>
<td>4000</td>
<td>617.8</td>
<td>586</td>
<td>6.886</td>
<td>623.4</td>
</tr>
<tr>
<td>5000</td>
<td>615.8</td>
<td>586</td>
<td>6.540</td>
<td>619.8</td>
</tr>
<tr>
<td>6000</td>
<td>615.8</td>
<td>586</td>
<td>6.540</td>
<td>617.2</td>
</tr>
<tr>
<td>7000</td>
<td>614.0</td>
<td>586</td>
<td>6.228</td>
<td>616.8</td>
</tr>
<tr>
<td>8000</td>
<td>613.4</td>
<td>586</td>
<td>6.125</td>
<td>614.6</td>
</tr>
<tr>
<td>9000</td>
<td>611.4</td>
<td>586</td>
<td>5.779</td>
<td>614.0</td>
</tr>
<tr>
<td>10000</td>
<td>611.4</td>
<td>586</td>
<td>5.779</td>
<td>612.6</td>
</tr>
</tbody>
</table>

Table 2: Solving abilities for 12-elements QAP. (αc=3.6)

<table>
<thead>
<tr>
<th>Iteration</th>
<th>switching chaos</th>
<th>switching random</th>
<th>switching torus</th>
<th>Cubic Map</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ave</td>
<td>Min</td>
<td>error</td>
<td>Ave</td>
<td>Min</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------</td>
<td>-----------</td>
</tr>
<tr>
<td>1000</td>
<td>635.8</td>
<td>612</td>
<td>10.000</td>
<td>637.0</td>
</tr>
<tr>
<td>2000</td>
<td>633.6</td>
<td>612</td>
<td>9.619</td>
<td>627.4</td>
</tr>
<tr>
<td>3000</td>
<td>623.8</td>
<td>598</td>
<td>7.924</td>
<td>624.8</td>
</tr>
<tr>
<td>4000</td>
<td>621.4</td>
<td>598</td>
<td>7.509</td>
<td>617.8</td>
</tr>
<tr>
<td>5000</td>
<td>619.2</td>
<td>598</td>
<td>7.128</td>
<td>613.6</td>
</tr>
<tr>
<td>6000</td>
<td>614.6</td>
<td>598</td>
<td>6.332</td>
<td>613.6</td>
</tr>
<tr>
<td>7000</td>
<td>610.6</td>
<td>598</td>
<td>5.640</td>
<td>613.4</td>
</tr>
<tr>
<td>8000</td>
<td>609.4</td>
<td>598</td>
<td>5.433</td>
<td>613.4</td>
</tr>
<tr>
<td>9000</td>
<td>609.4</td>
<td>598</td>
<td>5.433</td>
<td>611.8</td>
</tr>
<tr>
<td>10000</td>
<td>608.4</td>
<td>598</td>
<td>5.259</td>
<td>611.8</td>
</tr>
</tbody>
</table>
Abstract- The purpose of this paper is to build an optimized quinary multiplier by the help of the logic oriented neural network [1] (LOGO-NN) and by the use of the mixed radices of binary / quinary. It has been found [2] that mixed radices (binary / quaternary) provide a convenient way to analyze, synthesize and minimize the quaternary logic functions. Also, it has proven that the mixed radices LOGO-NN enable us to reduce the number of elements and interconnections. Here we will try to make mixing of radices between binary and quinary but by using quaternary as an intermediate stage.

1. Introduction
Recent advances in neural sciences and microelectronic technologies have greatly increased the challenges of development of high speed computational capability and efficiency of complicated engineering tasks. In this paper, a new approach of neural networks to implement a quinary arithmetic multiplier model has been proposed. The LOGO-NN is able to perform several independent computations in parallel [1] by a single network. The multiple-valued logic LOGO-NN [2, 7] provide powerful computational capabilities for large quantity of data. The new LOGO-NN system is accompanied by mathematical tools which allow us to analyze and synthesize any logic models in a very simple systematic way. The LOGO-NN is proposed in order to form a complete system that can realize any multiple-valued logic function [3].

2. Neuron model
The LOGO-NNs are composed of one neuron type and all synapse’s weights between neurons are natural integers. These two characteristics make LOGO-NNs useful, simple to design and more realistic in comparison with that of [4]. Since the Galois field algebra provides a convenient way to specify the structure of binary [5], ternary [6] and quaternary [2], the LOGO-NN operators of Galois field along with the logic constants form a finite field.

A LOGO-NN structure of k-valued logic is defined as:

$$\text{NNQ} = (G, \text{GF} (k), f (Z))$$ (1)

Where,
- \(G\): Finite directed graph, and is defined as
  $$G = (N, L, W)$$ (2)

Where,
- \(N\): Set of nodes (processing element or neurons),
- \(L\): Set of links (connections),
- \(W\): Set of synapse’s weights.

- \(\text{GF} (k)\): Galois field of k elements.
  $$\text{GF} (k) = \{0, 1, 2, \ldots, k - 1\}$$ (3)
  $$k \geq 2$$ (4)

- \(f (Z)\): Output signal of processing elements (neuron)
  $$f (Z) = \begin{cases} Z & \text{if } Z > 0 \\ 0 & \text{if } Z \leq 0 \end{cases}$$ (5)

$$Z = \sum_{i=0}^{n} x_i w_i - \theta$$ (6)

Figure 1: Processing element

Where,
- \(x_i\): Input signals,
- \(x_i \in \text{GF}(k) = \{0, 1, 2, \ldots, K - 1\}\),
- \(w_i\): Multiplicative coefficient (weight) for \(x_i\),
- \(i = 0, 1, \ldots, n\),
- \(\theta\): Threshold of the processing element,
- \(w_i, \theta \in \{\ldots, -2, -1, 0, 1, 2, \ldots\}\)

Figure 2: Linear transfer function.

3. Galois field of 2 elements
Any binary logic function can be represented by the familiar Galois field structures [2]. The flexibility of this modular algebra demonstrated its suitability for the applications of LOGO-NN. The Galois filed of 2-element is defined as \(K = 2\), then \(\text{GF} (2) = \{0, 1\}\). Where GF(2) is defined by addition \(\oplus\) and multiplication \(\cdot\) Functions as given in the table 1. and as shown in figure 3.
Table 1: ⊕ and • Functions of GF(2).

<table>
<thead>
<tr>
<th>⊕</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>•</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 3: LOGO-NN of GF(2) functions.

4. Basic Binary LOGO neural networks

4.1 Complement function

The complement function is defined by Equation 7 and Table 2, where its LOGO-NN operator is designed as shown in Fig. 4.

\[ x = 1 - \bar{x} \]  
(7)

Table 2: Complement Functions

<table>
<thead>
<tr>
<th>x</th>
<th>(\bar{x})</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 4: LOGO-NN of complement function

4.2 GF (2) multiplication of n-input signals

The LOGO-NN of GF(2) multiplication of n-input signals (Fig. 5) is designed for:

\[ \omega_0 = \omega_1 = \omega_2 = \ldots = \omega_n = 1 \quad \text{and} \quad 0 = n \]

Then:

\[ f(Z) = x_0 \cdot x_1 \cdot x_2 \cdot \ldots \cdot x_n \]  
(8)

Figure 5: LOGO-NN for GF(2) multiplication of n-input.

4.3 OR function

The OR function is defined as given in Table 3 and as shown in figure 6.

Table 3: OR Functions

<table>
<thead>
<tr>
<th>+</th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 6: LOGO-NN of OR function

5. Minimization rule of LOGO Neural Networks

The LOGO neural networks of Fig. 7 shows a simple example for a reduction rule that can be used to minimize LOGO-NN of the expression ( \( f = x \cdot y \cdot z \) )

6. Quinary Logic Multiplier

The quinary logic multiplication process of two input variables is defined as given in the table 4, where \( M \) is the multiplier ( \( M = X \cdot Y \) ) and \( C \) is the carry of \( M \).

Table 4: Quinary Multiplication process

<table>
<thead>
<tr>
<th>Y</th>
<th>X</th>
<th>M</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
X, Y and M belong to quinary set \{0,1,2,3,4\}. To represent these variables in binary, we need three bits for each. Then we will obtain 48 different binary combinations between X and Y where 23 of them will be unused.

Let: \(Y=(Y_3,Y_2,Y_1), X=(X_3,X_2,X_1), M=(M_3,M_2,M_1)\) while \(C=(c_2,c_1)\)

To simplify the matter, we will try to represent the variables by two bits only then the problem will become similar to the quaternary issue [2] where we will get only 16 binary combinations and that will lead to minimize the expressions of the functions of \(M\) and \(C\).

For this purpose, we will apply the following:

- We have noticed from table 4 that \(M=C=0\) when \(X=0\) or \(Y=0\) then let us omit and remove these cases from table 4. The values of \(X, Y\) and \(M\) now belong to the set \{1,2,3,4\}.

- After removing the cases of \(X=0\) or \(Y=0\), let us substitute 1 from all digits of \(X, Y\) and \(M\). The set of numbers becomes the same of quaternary one \{0,1,2,3\}.

- After substituting 1, let us represent all numbers 0,1,2 and 3 by two bits binary vectors where 0=(0,0), 1=(0,1), 2=(1,0) and 3=(1,1). So with applying the following functions, the table 4 becomes as shown on table 5.

\[\begin{align*}
y &= Y-1=(y_2,y_1) \\
x &= X-1=(x_2,x_1) \\
m &= M-1=(m_2,m_1) \\
C &= (c_2,c_1)
\end{align*}\]

Table 5: Substitution 1 from X,Y and M for X≠0, Y≠0

<table>
<thead>
<tr>
<th>y2,y1</th>
<th>x2,x1</th>
<th>m2,m1</th>
<th>c2,c1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,0</td>
<td>0,0</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>0,1</td>
<td>0,1</td>
<td>1,0</td>
<td>1,0</td>
</tr>
<tr>
<td>0,0</td>
<td>0,0</td>
<td>0,1</td>
<td>0,0</td>
</tr>
<tr>
<td>0,1</td>
<td>0,1</td>
<td>1,1</td>
<td>1,1</td>
</tr>
<tr>
<td>0,0</td>
<td>0,0</td>
<td>0,1</td>
<td>0,1</td>
</tr>
<tr>
<td>0,0</td>
<td>1,0</td>
<td>1,0</td>
<td>0,0</td>
</tr>
<tr>
<td>0,1</td>
<td>0,1</td>
<td>1,1</td>
<td>0,0</td>
</tr>
<tr>
<td>0,1</td>
<td>1,0</td>
<td>0,0</td>
<td>0,1</td>
</tr>
<tr>
<td>0,1</td>
<td>1,1</td>
<td>0,1</td>
<td>0,1</td>
</tr>
<tr>
<td>1,0</td>
<td>0,0</td>
<td>0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>1,0</td>
<td>0,1</td>
<td>0,0</td>
<td>0,1</td>
</tr>
<tr>
<td>1,1</td>
<td>0,1</td>
<td>0,0</td>
<td>1,0</td>
</tr>
<tr>
<td>1,1</td>
<td>0,1</td>
<td>0,1</td>
<td>1,0</td>
</tr>
<tr>
<td>1,1</td>
<td>1,0</td>
<td>0,0</td>
<td>1,0</td>
</tr>
<tr>
<td>1,1</td>
<td>1,1</td>
<td>0,0</td>
<td>1,1</td>
</tr>
</tbody>
</table>

Now we have a traditional functions (m1,m2,c1 and c2) with four variables. Using karnaugh map [5] to minimize these functions, we obtain that:

\[\begin{align*}
m_1 &= x_1.y_1.y_2 + x_2.y_1.y_2 + x_1.y_1.y_2 + x_2.y_1.y_2 \quad (13) \\
m_2 &= x_2.y_1.y_2 + x_1.y_1.y_2 + x_2.y_1.y_2 + x_1.y_1.y_2 \quad (14) \\
c_1 &= x_2.y_1.y_2 + x_1.y_1.y_2 + x_1.x_2.y_2 + x_1.x_2.y_1.y_2 \quad (15) \\
c_2 &= x_2.y_1.y_2 + x_1.x_2.y_2 + x_1.x_2.y_2 \quad (16)
\end{align*}\]

But we have to find \(M(M_3,M_2,M_1)\) where \(M=(m+1)\) Then the table 5 becomes as in table 6.

Table 6: Addition of 1 for m (M=m+1)

<table>
<thead>
<tr>
<th>y2,y1</th>
<th>x2,x1</th>
<th>M3,M2,M1</th>
<th>c2,c1</th>
</tr>
</thead>
<tbody>
<tr>
<td>0,0</td>
<td>0,0</td>
<td>0,0,1</td>
<td>0,0</td>
</tr>
<tr>
<td>0,0</td>
<td>0,1</td>
<td>0,1,0</td>
<td>0,0</td>
</tr>
<tr>
<td>0,0</td>
<td>1,0</td>
<td>0,1,1</td>
<td>0,0</td>
</tr>
<tr>
<td>0,0</td>
<td>1,1</td>
<td>1,0,0</td>
<td>0,0</td>
</tr>
<tr>
<td>0,1</td>
<td>0,0</td>
<td>0,1,0</td>
<td>0,1</td>
</tr>
<tr>
<td>0,1</td>
<td>0,1</td>
<td>1,1,0</td>
<td>0,0</td>
</tr>
<tr>
<td>0,1</td>
<td>1,0</td>
<td>0,0,1</td>
<td>0,1</td>
</tr>
<tr>
<td>0,1</td>
<td>1,1</td>
<td>1,0,1</td>
<td>0,1</td>
</tr>
<tr>
<td>1,0</td>
<td>0,0</td>
<td>0,0,0</td>
<td>0,1</td>
</tr>
<tr>
<td>1,0</td>
<td>0,1</td>
<td>0,0,1</td>
<td>1,0</td>
</tr>
<tr>
<td>1,0</td>
<td>1,0</td>
<td>1,0,0</td>
<td>0,1</td>
</tr>
<tr>
<td>1,0</td>
<td>1,1</td>
<td>1,0,1</td>
<td>0,1</td>
</tr>
<tr>
<td>1,1</td>
<td>0,0</td>
<td>0,1,0</td>
<td>1,0</td>
</tr>
<tr>
<td>1,1</td>
<td>0,1</td>
<td>0,1,1</td>
<td>0,1</td>
</tr>
<tr>
<td>1,1</td>
<td>1,0</td>
<td>0,0,1</td>
<td>1,0</td>
</tr>
<tr>
<td>1,1</td>
<td>1,1</td>
<td>0,0,1</td>
<td>1,1</td>
</tr>
</tbody>
</table>

From table 5 and 6, we conclude the equations of M:

\[\begin{align*}
M_1 &= \overline{m_1} \quad (17) \\
M_2 &= m_1 \oplus m_2 \quad (18) \\
M_3 &= (m_1,m_2) \quad (19)
\end{align*}\]

Figure 8 shows the LOGO-NN for equations of the carry C and the multiplier M (15,16,17,18 and 19). But in the cases where X=0 OR Y=0 all equations of M and C should equal to zero. Then the final equations of M should be anded by X.Y where:

\[X.Y = \begin{cases} 0 & \text{if } X=0 \text{ or } Y=0 \\ 1 & \text{otherwise} \end{cases}\]

\[\begin{align*}
M_{11} &= \overline{m_1}.X.Y \quad (21) \\
M_{22} &= (m_1,m_2).X.Y \quad (22) \\
M_{33} &= (m_1.m_2).X.Y \quad (23) \\
c_1 &= (x_2.y_1.y_2 + x_1.y_1.y_2 + x_1.x_2.y_2 + x_1.x_2.y_1.y_2).X.Y \quad (24) \\
c_2 &= (x_2.y_1.y_2 + x_1.x_2.y_2).X.Y \quad (25)
\end{align*}\]
7. Conclusion

This paper presented the development of a new technique concerning the quinary multiplier LOGO-NN through the use of mixed radices (binary / quinary) in order to simplify, and minimize as possible the used elements and to increase the performance of quinary multiplier to the maximum. The advantages of mixing the radices are the simplicity of binary design and the availability of binary components. The algorithm here make possible the representation of the quinary as binary with two variables only while in general it needs to be represented with three bits for each quinary digit (quits). For evaluation, the advantages of the proposed quinary LOGO neural network multiplier in comparison with other techniques [1,2,4] are the following:

1. The proposed multiplier is composed of one neuron type, while that of [4] is composed of three neuron types.
2. The proposed multiplier performs the complete operation in a single LOGO neural network, while that of [4] is decomposed into three sub-circuits which are controlled by an individual circuit.
3. Integers are representing the synapse’s weights, and thresholds of neurons while that of reference [1] are non integers numbers.
4. The interconnection complexity is decreased comparing with Back propagation neural networks algorithm [1].
5. The proposed LOGO has linear transfer function, while in [1] has sigmoid function (non linear). Therefore, the computation process in our proposal will be faster and has the ability to implement several independent computations in parallel by a single network.

References

Abstract

In this paper we propose a path searching method based on the local Winner Take All (WTA) circuits. The WTA is constructed by corporative and competitive networks to find the maximum local current on each node connected to variable resistance, and it is used for planning the driving path of a robot. The local current comparison method is deeply related to the node analysis for Cellular Neural Network (CNN). In our method, we used the $m \times n$ size of the lattice net. The $m \times n$ size of the lattice net was able to express various maps. By putting WTA on each node, we could find the path at the short time.

1. Introduction

Recently theories and applications about neuro-computing are researched actively. However it is difficult to integrate those conventional neural networks onto a silicon chip because of their structural complexity. So the method called retinal information processing has been remarkable for its easiness of fabrication on a chip and for its concurrency.

For example, the cellular neural network (CNN) [1][2] is a locally interconnected, large-scale nonlinear analog neural network. In the principle of the CNNs information is transmitted only to neighboring neurons, and composition of the circuits of its neurons is not so complex comparatively, so they are fit for integration, and we can apply parallel processing to them.

The WTA is deeply related to the node analysis for CNN. It is commonly used in a wide variety of applications. One of the most used current-mode WTA designs is proposed by Lazzaro et al. [3]. We used WTA for finding the maximum local current on each node connected to variable resistance.

In this paper, we propose a novel solution of the path planning problem based on local current comparison method (LCC method) using an analog variable resistive network and competitive network. Each resistance is corresponding to the slope or length of a road. It is based on retinal information processing and analog dynamics, and can be speeded up by realizing with the hardware.

In this paper, the $m \times n$ size of the lattice net was realized by extending the $n \times n$ lattice. By generalizing the size of lattice, we came to be able to make the lattice net correspond to various maps. We have applied this novel method to multi-robot driving control and will show some simulation results of it.

2. Winner Take All

The competitive network by which the branch having maximum local current is selected at each node is illustrated in figure 1. The network is constructed of four elements $E_1$, $E_2$, $E_3$, and $E_4$, which have self-feedback loop with the weight $w_i$, and one suppressive

![Figure 1: The competitive network to select the branch of maximum local current.](image)
element \( I \). The suppressive element \( I \) gets the output value of elements \( E_i \) with the weight 1 and its output is connected to \( E_i \) with the weight \(-w_2\).

Each element works in continuous time \( t \) and circuit equations of the element \( E_i \) is defined as follows:

\[
\frac{du_i(t)}{dt} = -u_i(t) + w_1 f(u_i(t)) - w_2 g(v(t)) - h_1 + s_i,
\]

(1)

\[
\tau \frac{dv(t)}{dt} = -v(t) + \sum_{i=1}^{n} f(u_i(t)) - h_2,
\]

(2)

where \( h_1, h_2, w_1, w_2, \) and \( \tau \) are constants, \( f(u) \) is the output function of the exciting element, and \( g(v) \) is the suppressive element respectively. Functions \( f(u) \) and \( g(v) \) are given by:

\[
f(u) = \frac{1}{1 + \exp(-u/T)},
\]

(3)

\[
g(v) = \begin{cases} v & v > 0 \\ 0 & v < 0. \end{cases}
\]

(4)

In our system, \( s_1, s_2, s_3, \) and \( s_4 \) are got as input value and four elements \( E_1, E_2, E_3 \) and \( E_4 \) compete with each other.

This network is arranged at each node respectively, and the local current values of the four branches connected to the node are put in elements \( E_i \) as input \( s_i \). In the equilibrium state, only one element in which most large value was put become active, and it indicates the branch of maximum local current.

The convergence condition to the state of the stability equilibrium in competitive network is as follows:

\[
h_1 = s_{\text{max}} > 0
\]

\[
0 < h_2 < 1
\]

\[
w_2 > (s_{\text{max}} - h_1)/(1 - h_2)
\]

\[
w_2(2 - h_2) + h_1 - s_{\text{max}} > w_1 > w_2(1 - h_2) - h_1
\]

By the expression above, each constant of the network is provided as follows.

\[
w_1 = 1.6, \ h_1 = 0.005, \ \tau = 0.5,
\]

\[
w_2 = 2.0, \ h_2 = 0.45.
\]

Using the constant, a multi-dimensional differential equation of equation (2) and (3) is solved by using the Runge-Kutta method, and state \( u_i(t) \) of the element and \( v(t) \) are obtained.

Figure 2 shows the output result of excitatory elements \( E_1, E_2, E_3 \) and \( E_4 \) and inhibitory element \( I \). The cooperative and competitive state is from 0 to 6 second and the steady state is after 6 second. The largest input value \( E_4 \) wins, and the others are controlled.

3. Application to path searching by WTA

The path can be replaced with the resistance network. The node is replaced as a intersection. The resistance is replaced with consideration of the distance and the crowded condition to the following intersection. The path at the shortest time can be found by making the robot advance toward the maximum current. The WTA enables to find the maximum current.

In our method, the size of the lattice was adjusted to \( m \times n \). First, we made the size \( n \times n \) of the lattice net. In any size lattice net, we made the program that the voltage of each node was requested by substituting
the value of each resistance. The robot advances to the maximum current value. Therefore, if there is a path that the robot cannot pass, it only has to reduce the current value very much. In fact, it only has to enlarge resistance value very much (100000 Ω). As a result, the m × n lattice was realized by the n × n lattice.

The robot is made a current source and the robot aims at the terminal point from the start point. Whenever the robot moved, the node voltage of each node is requested.

By putting WTA on each node, the maximum current is detected by the current value from the node that is now to the following each node as shown in figure 3. The moving direction of the robot is as follows:

- case 1 ) \( i_1 > i_2 > i_3 > i_4 \)
  
  The moving direction of the robot is south where the maximum current value \( i_1 \) flows.

- case 2 ) \( i_1 = i_2 = i_3 = i_4 \)
  
  The moving direction of the robot is south, according to the predetermined priority level. If there are same maximum current values, the direction of the priority level is determined by \( i_1 > i_2 > i_3 > i_4 \).

As a result, the neighboring direction of the movement can be found. When it reaches the terminal by this repetition, the program is ended.

4. Results of Simulation

![Figure 4: The map around Kyoto station.](image)

We had simulation of the path planning for multi-robot driving and show the simulation result. We used the map around Kyoto station. As show in figure4, the start point is Kyoto Station, and the destination point is Ginkakuji temple. The road that was able to pass was shown in the heavy line. • is the intersection. We assumed that the heavy black line was the double crowded as much as the heavy gray line. The speed of the robot is constant. Therefore we can assumed that fraction of the distance was value of resistance. And the part of heavy black line made double value of resistance as show in figure5.

![Figure 5: The lattice net about Kyoto.](image)

By using our program, the path at the shortest time can be found. The simulation result was as show in figure6 and 7.

5. Conclusion

In this paper, we proposed the path searching using WTA method in which the branch of the maximum local current is selected by the competitive network, and applied it to the planning the driving path of the multi-robot. Using our method, the shortest route to the terminal is automatically detected, and it reaches the destination.

The local current comparison method is deeply related to node analysis in neighbor. And this method suits to be constructed as the hardware on the analog-digital hybrid chip, and it is expected that this method can be speeded up by realizing with the hardware which consists of analog resistive network and competitive networks. So, we will be able to apply this method for real time path searching and it can be
expected realization of the system which can generate plans for robots driving comparatively fast in real-time.

Acknowledgment

This research is supported by the fund of Open Research Center Project from MEXT of Japanese Government (2002-2006)

References

Identifying Returns Distribution by Using Mixture Distribution Optimized by Genetic Algorithm and Its applications

Kangrong Tan† and Shozo Tokinaga‡

†Faculty of Economics, Kurume University
1635 Miimachi, Kurume City, FUKUOKA, JAPAN
‡Graduate school of Economics, Kyushu University
6-19-1, Hakozaki, FUKUOKA, JAPAN
Email: tan_kouyuur@kurume-u.ac.jp, tokinaga@en.kyushu-u.ac.jp

Abstract—The main research of this paper is to propose a heuristic method to identify returns distribution more accurately than well-assumed conventional distributions, such as, normal, lognormal, or non-normal stable distribution. Though, these assumptions have their merits, many empirical researches have shown they are inconsistent with the business realities, since either one cannot catch excess kurtosis and fat tails behavior, or has infinite moments. Serious bias could be introduced when applied to risk management, i.e., evaluating the Value at Risk (VaR). We propose to identify returns distribution by using a mixture distribution constructed by weighted sum of some radical distributions whose parameters and weights are to be optimized by the Genetic Algorithm (GA).

1. Introduction

It is very important to identify asset returns distribution accurately since crucial statistical information can be obtained from the identified distribution, and can be applied to risk management, such as, evaluating VaR. An improperly selected distribution would provide wrong information and even lead to serious biases. So far, three distributional assumptions, namely, normal, lognormal, and non-normal stable distribution (for simplicity, we just call it stable distribution in this rest of the paper), have been proposed, and widely applied. But, many empirical analyses have shown these assumptions are in conflict with business realities, since normal or lognormal one cannot catch the characteristics of excess kurtosis or fat tails behavior, and stable distribution has infinite moments. It could introduce serious biases in some cases if they are simply applied without any consideration [1] [2] [5].

In this paper we propose to identify returns distribution more accurately than any above-mentioned assumptions by using mixture distribution, which is constructed as a weighted sum of a set of radical distributions, and the distributional parameters of the radical distributions and weights are to be optimized by GA. Through numerical experiments, a mixture distribution constructed by our proposed method shows it can overcome those demerits contained in the above-mentioned assumptions. The rest of this paper is organized as follows. In section 2, we summarize the merits and the demerits of the above-mentioned distributional assumptions. In section 3, we show how to construct mixture distribution based upon a set of radical distributions and how to optimize the parameters and weights by GA. In section 4, we present some applications and their numerical results. Finally in section 5, we provide some concluding remarks.

2. Merits and demerits of distributional assumptions

So far, normal, lognormal, and stable distribution have been well assumed and well applied in dealing with returns distribution. The common return definitions are, seen from Table-1, 1) one-period simple return, 2) multi-period simple return 3) one-period compounded return, and 4) multi-period compounded return, where \( p_t \) is asset price at time \( t \). Obviously, simple multi-period return can be regarded as product of each one-period return. And compounded multi-period returns can be regarded as sum of each single-period return. That is a significant difference between the two ones. It leads to be completely different distributions on multi-period returns. Some merits and demerits of these distributional assumptions are discussed and summarized as follows [1] [2] [4] [7] [8] [10].

<table>
<thead>
<tr>
<th>Return Definition</th>
<th>Expression</th>
</tr>
</thead>
<tbody>
<tr>
<td>One-period simple</td>
<td>( R_t = \frac{p_t}{p_{t-1}} )</td>
</tr>
<tr>
<td>Multi-period simple</td>
<td>( 1 + R_t[K] = \frac{p_t}{p_{t-1}} )</td>
</tr>
<tr>
<td></td>
<td>= ( \prod_{j=1}^{k}(1 + R_{t+j}) )</td>
</tr>
<tr>
<td></td>
<td>( R_t[k] = \frac{p_t}{p_{t-k}} )</td>
</tr>
<tr>
<td>One-period compounded</td>
<td>( r_t = ln(1 + R_t) = ln \frac{p_t}{p_{t-1}} )</td>
</tr>
<tr>
<td>Multi-period compounded</td>
<td>( r_t[k] = ln(1 + R_t[k]) )</td>
</tr>
<tr>
<td></td>
<td>= ln(1 + R_t) + ... + ln(1 + R_{t-k+1}) + ... + r_{t-k+1} )</td>
</tr>
</tbody>
</table>

Table 1: Return definitions
1) Normal distribution

Usually $R_t$ is assumed as an i.i.d normal process.

**Merit and demerit**

The merit of such a setting is that it makes returns to be tackled very easily. But, the demerits are, a) simple returns’ low bound is -1, but there’s no low bound in normal distribution, b) if $R_t$ is normal then $R_{t[K]}$ is not a normal one, c) it is not supported by many empirical results since it cannot catch excess kurtosis.

2) Lognormal distribution

Usually assumed that $r_t$ is an i.i.d normal process.

**Merit and demerit**

The merits are, a) there is no low bound existing in such a setting, b) it allows that $r_t[k] = ln(1 + R_t) + \ldots + ln(1 + R_{t+k})$ is also normally distributed. But, it is not supported by many empirical analyses, since it cannot catch excess kurtosis and fat tail behavior either.

3) Non-normal stable distribution

**Merit and demerit**

It allows the sum of returns still be stable distribution. And it can catch excess kurtosis and heavy tails well in many cases. But, the problem is that non-normal stable distribution has infinite moments. The estimates of variance and kurtosis tend to be larger and larger and not to converge even if sample size increases. It doesn’t match the reality of the financial market. Furthermore, it is complicated when it is applied to risk management.

3. Identifying returns distribution

3.1. Mixture distribution

Here, we assume that returns are always calculated by $r_t = lnP_t - lnP_{t-1}$, and then standardized to have mean 0 and variance 1. We propose to identify returns distribution as a weighed sum of radical distributions, especially we choose normals with different means and variances, or Student-ts with different d.f.s as radical distributions. Namely, let

$$r_t \sim \sum_{i=1}^{n} \beta_i \phi_i, \quad \sum_{i=1}^{n} \beta_i = 1 \quad (1)$$

where $\phi_i$ is a radical distribution (normal or student-t). Why normal distributions are chosen as radical distributions is that it allows mixture distribution do catch the characteristics of excess kurtosis and fat tails behavior much better than the conventional normal distribution. Such a mixture distribution has many good statistical properties that prevail over those demerits contained in the conventional assumptions, i.e., it has finite moments which stable distribution doesn’t have. The p.d.fs of N(0,1), Studen-t ($\nu=2.5$), Mixture distribution(S(1- $\beta$) N(0,1) + $\beta$ N(0,25)), Cauchy ($\frac{1}{\pi(1+x^2)}$), and N(0,25), where $\beta=0.05$, and 0.25 are shown in Figure-1. Seen from Figure 1, the kurtosis and tails of a mixture distribution can position themselves between N(0,1) and Cauchy distribution, adjusted by changing parameter $\beta$ and $\sigma_i$s. Thus, it can catch the characteristics of returns with finite moments. Benefit in this case is that the combination of normals in mixture distribution will be easily applied in risk management, i.e., evaluating the VaR. And scaled Student-t can more accurately catch heavy tails in some cases combined with normals.

3.2. Optimization of parameters and weights

Usually it is hard to estimate the distributional parameters and weights in mixture distribution. Here we suggest to apply GA to it. GA is known as one of the optimization methods, which converges hardly in a local optimal solution while searching a global one. It has been widely applied in many problems ranging from scientific studies to social studies [3][9].

The basics of our GA are as follows. Step 1) Generate individuals by random numbers as the first generation with certain population, each individual represents a set of parameters and weights. Step 2) Evaluate each individual by predetermined fitness function to get its fitness value. Step 3) Select two individuals with higher fitness through generations (crossover, mutation operation) to them to reproduce new individuals (offsprings) as the next generation. Step 4) Reevaluate the fitness of each individual of the new generation. If the results meet the Termination Conditions (i.e., repeating times, or error range), then GA terminates, otherwise it goes back to step 3).

And fitness function for $j$th individual is defined by

$$\text{Fitness}_j = (1 - \frac{\sum_{i} e_{ij}^2}{\sum_{j} \sum_{i} e_{ij}^2})/(n - 1) \quad (2)$$

Figure 1: Up: Plots of p.d.fs: top to bottom: N(0,1), Mixture, t-2.5, Cauchy, N(0,25). Down: Plots of p.d.fs: top to bottom: N(0,1), t-2.5, Mixture, Cauchy, N(0,25)(right)
where $e^2_i = (f_i - f'_i)^2$, $f_i$s and $f'_i$s are observations and estimates respectively.

4. Applications

In this section, we present some numerical applications. In application 1, we apply our proposed method to daily stock price data set A (01/04/1985 ~ 25/06/2001), B (01/04/1983 ~ 21/09/2001), and C (01/01/1962 ~ 08/12/2004, 9-day interval) to obtain mixture distributions. We compare the numerical results from mixture distributions with the ones from the conventional normals and find our results are more accurate than the normals’. In application 2), we apply our method to evaluating the VaR. We also find mixture distributions provide more accurate estimates of the VaR than the normals do as well.

<table>
<thead>
<tr>
<th>Stock A</th>
<th>Stock B</th>
<th>Stock C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>1.136E-5</td>
<td>4.693E-05</td>
</tr>
<tr>
<td>Median</td>
<td>1.325E-4</td>
<td>7.107E-05</td>
</tr>
<tr>
<td>S.D</td>
<td>5.941E-3</td>
<td>0.0050128</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>9.330635</td>
<td>12.662960</td>
</tr>
<tr>
<td>Min</td>
<td>-0.12732</td>
<td>-0.070071</td>
</tr>
<tr>
<td>Max</td>
<td>0.053984</td>
<td>0.031477</td>
</tr>
<tr>
<td>Size</td>
<td>4189</td>
<td>4825</td>
</tr>
</tbody>
</table>

Table 2: Descriptive statistics of returns of Stock A, B & C

![Figure 2: Up: Plots of p.d.fs; solidline: estimated mixture distributions; dashline: standard normal distributions; points: observations. Down: Plots of p.d.fs; solidline: estimated mixture distributions; dashline: standard normal distributions; points: observations.](image)

or highest fitness into next generation automatically. And selection strategy used here is the roulette strategy.

Results for Stock A & B are shown in Table 3. Each mixture distribution is estimated by GA as a weighed sum of two normals with same means($\mu_i$s) and different standard deviations $\sigma_1$s and $\sigma_2$s. Namely, $r_i \sim (1 - \beta_i)N(\mu_i, \sigma^2_1) + \beta_iN(\mu_i, \sigma^2_2)$.

Seen from Figure 2, clearly, mixture ones are much more closer to observations than the conventional normals. Mixture distributions catch excess kurtosis and fat tails behavior much better than normals. Plots of fitnesses and prediction errors of individuals of the last generation are shown in Figure 3. Clearly, fitnesses and prediction errors have converged after GA has been implemented 50 generations.

For Stock C, it is estimated by GA as a weighed sum of three normals and one Student-t, namely, $\beta_1N(0.107733965, 1.730427742^2) + \beta_2N(0.079310417, 1.213914633^2) + \beta_3N(-1.70856595, 1.720302582^2) + \beta_4t(\nu=25.24713135)$,

where the scaling weights are $\beta_1=0.11709362$, $\beta_2=0.20807926$, $\beta_3=0.0250332$, $\beta_4=0.64979392$, respectively. And $\sum_i \beta_i = 1.19687E-05$. And note, here, $\sum_i \beta_i = 1$ holds.

It means that mixture distribution can identify returns distribution accurately even if strong skewness and kurtosis exist.

**Application-2:** Evaluation of VaR

We apply our proposed method to portfolio management. Three portfolios, $p_1$, $p_2$, and $p_3$, are consisted of 5, 12 and 50 assets respectively. Ratios of the VaR based upon mix-

![Figure 2: Up: Plots of p.d.fs; solidline: estimated mixture distributions; dashline: standard normal distributions; points: observations. Down: Plots of p.d.fs; solidline: estimated mixture distributions; dashline: standard normal distributions; points: observations.](image)
5. Concluding remarks

In this paper, we have summarized and discussed the merits and demerits of the well-applied conventional returns distributions (normal, lognormal, and stable distribution). One problem of the conventional normal distribution is that it cannot catch excess kurtosis and fat tails behavior well. And for stable distribution, the problem is that it doesn’t have finite moments. In order to solve these above-mentioned problems, we have proposed and shown on how to identify the returns distribution by a mixture distribution, namely, modeling the returns distribution as a weighed sum of a set of radical distributions, normal or Student-t distributions, for example. And we also have proposed to apply Genetic Algorithm to optimize the distributional parameters and weights in mixture distribution. In our numerical applications, we have confirmed that mixture distribution is much better than the conventional normal distribution in capturing excess kurtosis and heavy tails of returns distribution with finite moments, and provides more accurate estimates in risk measurement and management.

References


Table 4: Ratios of VaR based mixture distributions and normal distributions

<table>
<thead>
<tr>
<th>Portfolio</th>
<th>α</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>5%</td>
<td>1.082415556</td>
</tr>
<tr>
<td>p2</td>
<td>5%</td>
<td>1.065149874</td>
</tr>
<tr>
<td>p3</td>
<td>5%</td>
<td>1.216961073</td>
</tr>
</tbody>
</table>

Figure 3: Plots of fitness and errors of last generation
A decision based model of addiction: explaining quitting behavior

C. Mocenni†, G. Montefrancesco‡ and S. Tiezzi♯

† Department of Information Engineering, Univ. of Siena, Via Roma 56, 53100 Siena (Italy)
‡ Centro Studi sulle Dipendenze Patologiche, Univ. of Siena, Policlinico Le Scotte, 53100 Siena (Italy)
♯ Department of Economics, Univ. of Siena, P.zza S. Francesco 7, 53100 Siena (Italy)

Email: mocenni@ing.unisi.it, montef@email.it, tiezzisilvia@unisi.it

Abstract—In this paper a simulation model of the dynamics of addictive behavior is proposed. Starting from Bernheim and Rangel (B&R), addiction is modeled as a progressive susceptibility to stochastic environmental cues and is strongly related to the characteristics of the user, the substance and the environment. The work is mainly focused on finding feasible mathematical formulation of the model functions and parameters for obtaining successful attempts to quit, a pattern of behavior observed in clinical practice. Appropriate nonlinear settings of the model produce quitting behavior even though the addicted continues operating either in cold or hot mode. Our approach can thus be considered a generalization of the B&R results, where the addicted foresees drug abuse only by exercising cognitive control and behaving in a cold mode.

1. Introduction

Addiction is currently defined as the consequence of repeated use of psychoactive drugs. It is characterized principally by a loss of control over drug seeking behavior, with harmful effects on the individual and a high probability of relapse even months or years after cessation of drug taking [1], [2], [3].

The main problem is to understand how this phenomenon “moves”, meaning how the various components of its multifactoriality (individual, substance and environment) can trigger the start, sustain recurrence or generate the frustrating relapse.

Recent (economic) theories of addiction can be loosely classified as variations or generalizations of the rational addiction model by Becker and Murphy [5]. Such generalizations allow for the presence of random shocks [6]; “projection bias” [7]; present-biased preferences and sophisticated or naive expectations [8]; “temptation” [9], where well-being depends not only upon the chosen action, but also upon the action not chosen.

The Bernheim and Rangel (B&R) [10] model instead regards addiction as a progressive susceptibility to stochastic environmental cues that can trigger mistaken usage. The point of departure of this theory from previous ones is its attempt to harmonize economic theory with evidence from psychology, neuroscience and clinical practice so as to explain the relationship between behavior and the characteristics of the user, substance and the environment mentioned above [(10), p.1559].

Neuroscience and clinical practice have indeed shown that addictive substances systematically interfere with the proper operation of a process used by the brain to forecast near term hedonic rewards (called Hedonic Forecasting Mechanism, HFM henceforth) and lead to strong impulses towards consumption that may interfere with higher cognitive control.

The B&R model can explain several patterns of addictive behavior, but there is one aspect left unexplained which is abrupt and sudden quitting. Clinical practice shows that this is not an infrequent pattern of behavior in long term addicts, but the reasons for it are still to be recovered.

The main purpose of this paper is to find behavioral patterns in the B&R model where abrupt cessation from consumption occurs. This result is achieved by modeling the payoff function with a nonlinear function. Performance analysis of the modeling stage is carried out by comparing the nonlinear settings with more simple linear choices in the simulation parameters and functions.

The paper is organized as follows: Section 2 provides a clinical description of the addictive phenomenon, while Section 3 reports the mathematical formulation of the B&R model. Section 4 shows the results of linear and nonlinear settings of the model’s functions and parameters. Finally, conclusions and future work are discussed in Section 5.

2. The clinic of addictive behavior

In human beings drugs produce an increase of dopamine concentration at target-cells’ receptor levels, as they stimulate the nigrostriatal (controlling motor coordination) and corticolimbic (controlling emotions and cognitive abilities) dopaminergic systems [11]. These cerebral systems have evolved not to entertain addictive substances, but to ensure the survival of the individual by controlling basic functions such as mating, consumption, searching for food and water, etc. Once these systems are engaged by natural rewards (food, or sex, for example) or by addictive substances [12] dopamine release in the nucleus accumbens and in other cerebral sites increases, causing specific affective states (for example, euphoria), that are powerful drivers and reinforce that behavior. The individual is thus induced to repeat such positive experiences (or avoid them when negative), precisely because he associates
the specific function to its hedonic (likeable) effects [13], [14], [15]. Addictive substances have an advantage over natural rewards: they produce a higher dopamine concentration by stimulating the system more powerfully and for longer periods [16]. Moreover, in the case of natural rewards, a habit develops after some time which reduces the importance of the experiential effect. In other words the quality and quantity of the gained pleasure diminishes. Addictive substances, however, act like powerful “novelties” activating each time, in a non-decremental way, dopaminergic transmission even after repeated use.

Chronic substance abuse induces profound alterations of the cerebral mechanisms just mentioned, and they “force”, in a way, the user to make compulsory and “wrong” choices, i.e. choices that diverge from preferences. In fact, drugs, by powerfully activating dopaminergic transmission, reinforce excessively the associated learning process, ending up by constraining the individual’s behavioral choices [17].

The pleasure following use, the excessive and rapid hedonic expectation induced by the HFM, the progressive failure of the frontal cortex to counterbalance with rational choices the more alluring offer of drugs, all portray a process that invariably regenerates itself and seems to have no end [13], [18]. Although drug addiction seems to lead to just one possible result, for still unclear reasons, often the patient stops participating in the ineluctable dynamics of his/her case and ceases to have this insatiable hunger and compulsion for the drug. This may happen as a consequence of psychological, social, pharmacological and individual interactions as well as all other stimuli found inside and around an individual (deterministic or even stochastic events).

In more general terms, one could say that the multifactoriality sustaining drug addiction, sometimes, ceases to offer those profits or conveniences considered up till then as indispensable.

3. The Bernheim-Rangel Behavioral Model of Addictive Consumption

The B&R model is based on the following premises: a) consumption among addicts is frequently a mistake; b) previous experience with an addictive good sensitizes an individual to environmental cues that trigger mistaken usage; c) awareness of sensitivity to cue-triggered mistakes produces attempts to manage the process with some degree of sophistication.

B&R consider a decision maker (DM) who can operate either in a cold (involving rationality) or hot mode (where decisions and preferences may diverge). He lives for an infinite number of discrete periods and, in each of them, makes two decisions in succession. First, he selects a “lifestyle” a from the set \( \{E,A,R\} \); then he allocates resources between an addictive substance \( (x \in [0, 1]) \) and a non addictive substance \( (e \geq 0) \). Each period is entered in cold mode and the DM chooses his lifestyle rationally. This choice, along with his history of use \( (s) \) determines the probability \( p_e^s \) with which he encounters cues that trigger the hot mode. The volume of substance related environmental cues, \( c(a, \omega) \), depends on the lifestyle and on an exogenous state of nature \( \omega \) drawn randomly from a state space \( \Omega \) according to some probability measure \( \mu \). The impulses from the HFM defeat cognitive control and place the DM in hot mode when their strength, measured by the function \( M(c, s, a, \omega) \), exceeds some threshold \( M^f \).

In state \( s \), the DM receives an immediate hedonic payoff \( w_s(e, x, a) = u(e) + v_s(x, a) \), where the utility derived from non-addictive consumption, \( u(e) \), is separable from utility derived from addictive consumption \(^1\), \( v_s(x, a) \), and where \( v_s(x, a) \equiv u_e^a + b^x \). Here \( u_e^a \) represents the baseline payoff associated with successful abstention in state \( s \) and activity \( a \) and \( b^x \) represents the marginal instantaneous benefit from use the individual receives in state \( s \) after taking activity \( a \) ([10], P. 1566). The DM discounts future hedonic payoffs at a constant rate \( \delta \). His choices in the cold mode correspond to the solution of a dynamic stochastic programming problem with a value function \( V_s(\theta) \) satisfying:

\[
V_s(\theta) = \max_{(a(x), c)} \{ u_e^a + \sigma_e^{x,s} b^x_c + \delta B \},
\]

where

\[
B = \left[ \left( 1 - \sigma_e^{s+1,s} \right) V_{\max (1, s-1)}(\theta + \sigma_e^{s+1,s} V_{\min (S, s+1)}(\theta)) \right].
\]

In equation (1) \( C \) is the set of decision states \( \{(E, 1), (E, 0), (A, 0), (R, 0)\} \), while \( \sigma_e^{s+1,s} \) represents the probability of consuming the substance in state \( x \) with contingent plan \( (a, x) \); \( \theta \) is a vector specifying all pertinent derivative parameters, i.e. it reflects the properties of the substance, the method of administration, the characteristics of the individual use and the public policy environment. Finally, \( S \) is the highest addictive state. There are \( S + 1 \) addictive states labelled \( s = 0, 1, ..., S \). Usage in state \( s \) leads to state \( \min(S, s + 1) \) in the next period. No use leads to state \( \max(1, s-1) \) from state \( s > 1 \) and to state \( 0 \) from state \( 0 \).

In this model, unsuccessful attempts to quit occur when there is an unanticipated or anticipated and sufficiently slow shift in parameters \( \theta = (p_s, u, b, \cdot) \), from \( \theta' \) to \( \theta'' \). Suppose that at state \( 0 \) the best choice for the DM is \( (E, 1) \) if \( \theta' \) prevails forever and the best choice at state \( 1 \) is \( ((E, 0), (A, 0), (R, 0)) \) for all \( s \) if \( \theta'' \) prevails forever. However, with \( p_e^s > 0 \) the attempt is unsuccessful when \( [(E, 0), (A, 0)] \) is chosen in state \( 1 \).

4. Comparative analysis and simulation results

In this section a comparative analysis of the model options is presented. The analysis is performed by applying

\(^1\)\( w_s \) must be increasing, unbounded, strictly concave and twice differentiable with bounded second derivative in the variable \( e \).
different choices on the mathematical formulation of the main processes accounted by the B&R addiction model described in Section 3.

In particular, according to the theoretical assumptions proposed in [10], linear and nonlinear settings of the model functions are pointed out in Table 1.

Table 1: Model formulation and parameter settings.

<table>
<thead>
<tr>
<th>Function</th>
<th>Linear settings</th>
<th>Nonlinear settings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exogenous cues, $\omega$</td>
<td>normally distr. random variable</td>
<td>normally distr. random variable</td>
</tr>
<tr>
<td>Volume $c(E,\omega)$</td>
<td>$\omega + E$</td>
<td>$\omega + E'$</td>
</tr>
<tr>
<td>Volume $c(A,\omega)$</td>
<td>$\omega + A$</td>
<td>$\omega + A'$</td>
</tr>
<tr>
<td>Volume $c(R,\omega)$</td>
<td>$\omega + \sqrt{R}$</td>
<td>$\omega + \sqrt{R'}$</td>
</tr>
<tr>
<td>HFM impulses $M(c,s,a,\omega)$</td>
<td>$s(a,\omega) + s$</td>
<td>$s(a,\omega) + \sqrt{s'}$</td>
</tr>
<tr>
<td>Hedonic payoff $w_i(r,1,E)$</td>
<td>$E + s + \sqrt{E'} - q$</td>
<td>$E + \sqrt{E'} - q$</td>
</tr>
<tr>
<td>Hedonic payoff $w_i(r,0,\omega)$</td>
<td>$A + s + \sqrt{A'}$</td>
<td>$A + \sqrt{A'}$</td>
</tr>
<tr>
<td>Hedonic payoff $w_i(r,0,A)$</td>
<td>$R + s + \sqrt{R'} - r$</td>
<td>$R + \sqrt{R'} - r$</td>
</tr>
<tr>
<td>Hedonic payoff $w_i(r,0,R)$</td>
<td>$R + s + \sqrt{R'}$</td>
<td>$R + \sqrt{R'}$</td>
</tr>
</tbody>
</table>

Lifestyles $E, A, R$, income $y$, rehabilitation costs $r$, substance price $q$ and activating threshold $M'$ are chosen constant.

Figure 1: Hedonic payoff $w_i$ in the linear settings.

Figures 1-2 show the simulation results when linear payoff functions are chosen: increasing values of the function $V$ in the simulation period correspond to an addictive behavior with a stable dynamics of 2 period consumption - 3 period avoidance states. Similar dynamics can be achieved by modifying, as an example, the substance costs $q$ (see Figure 2). The linear model settings may be unable to account for situations where any kind of quitting behavior is achieved; Figure 2 shows that the history of use is always increasing.

On the contrary, according to the clinic analysis of addictive behavior, many cases of abstention in drug consumption are observed. Furthermore, these situations are not only due to the coercitive consequences of substance abuse, such as public policies or enforcement costs. Subjective intentional abstention can also be a realistic scenario, due to endogenous modifications and spontaneous rehabilitation. Therefore a more complex setting up of the functions used in the model is proposed. In particular, a nonlinear relaxing payoff function is chosen, as described in the second column of Table 1. This choice is motivated by assuming that the instantaneous hedonic payoff of drug consumption can suffer, after a stiff increase in the beginning phase, negative feedbacks and reach asymptotic values (determined by the value of the parameter $h$), so that the DM can move into intentional abstention. This specification of the payoff function is also consistent with a decreasing natural rewards after some time of experimentation, as described in Section 2.

The simulated behavior corresponding to these model settings is shown in Figures 3-5. In particular, Figure 3 shows the cue triggering mechanism described in Section 3 and Figure 4 reports the nonlinear hedonic payoff function. Finally, in Figure 5 addiction ($h = 3$), recidivism ($h = 1.6$), successful attempt to quit ($h = 0.3$) and rehabilitation ($h = 0$) are simulated.
5. Conclusions

In this paper we have simulated the B&R addiction model in an attempt to achieve permanent and sudden abstinence from drug use, a consumption pattern often observed in clinical practice. This is obtained by choosing nonlinear functions describing the instantaneous payoffs evaluation. Our nonlinear setting of the model produces more general results than the B&R model in terms of the simulation results. In the latter, successful attempts to quit addiction depend on the assumption that the addicted exercises cognitive control and acts in a cold mode (p_a = 0) after an unanticipated or anticipated and sufficiently slow change in the relevant vector of parameters (see [10], p. 1571 for details). In our simulations, instead, successful attempts to quit may occur even when the addicted continues to operate either in cold or hot mode (p_a ≥ 0).

Future work will focus on testing model simulations with real patterns of DM’s history in order to obtain more reliable estimations of the model parameters.

References

TARGET AND SPIRAL WAVES IN CARDIAC CULTURES

O.I.Kanakov, G.V.Osipov† and C.K.Chan‡
†Radiophysical Dept., University of Nizhny Novgorod, Gagarin ave. 23, Nizhny Novgorod, Russia, 603950
‡Institute of Physics, Academia Sinica, Taipei, Taiwan
Email: okanakov@rf.unn.ru, osipov@rf.unn.ru, ckchan@gate.sinica.edu.tw

Abstract— We study collective phenomena in nonhomogeneous chains and lattices of coupled models of cardiac cells. Global and cluster synchronization, appearance of spiral and target waves were found.

1. Introduction

Modelling biological systems such as neuronal ensembles, kidney and cardiac tissues is one of the most rapidly developing fields of application of nonlinear dynamical methods nowadays. The efficiency of these methods is conditioned by the complex, though deterministic behavior of individual cells constituting the tissue.

In particular, cardiac cells exhibit the properties of either excitable or oscillatory systems. The former case is observed in working myocardium, and the latter is found in natural cardiac pacemakers (sinoatrial and atrioventricular nodes, Purkinje fibers). Normal heart activity is controlled by waves of excitation generated in the sinoatrial node and propagating through the conducting system and working myocardium. Deviations from the normal regime (arrhythmias) are often associated with pathological types of wave dynamics in the cardiac tissue. They include spiral waves and spiral chaos (the latter manifests itself in heart fibrillation).

In the present paper we report of a series of numerical experiments with cardiac cell culture models. We focus the main attention on the transition from incoherent behavior of uncoupled cells to global synchronization in ensembles of strongly coupled cells, when the coupling coefficient is increased from zero. This corresponds to the increase of the number of gap-junctions in the culture.

We show, that this transition occurs via cluster synchronization regimes in one-dimensional models and complicated spatio-temporal activity in two-dimensional models. These dynamical effects arise from spatial discreteness and inhomogeneity of the model.

Similar experiments in-vitro were reported in [1]. According to [1], after approximately 24 hours of culture time, irregular spontaneous activity arises in the culture, further it organizes itself into several pacemakers emitting target waves. These pacemakers are subsequently destroyed, and spiral wave activity sets in; the number of spiral cores is changing in time [1].

2. The model

2.1. Excitable cells

We use the Luo-Rudy phase I model [2] to define the dynamics of a single cell. This model describes the dynamics of excitable cardiac cells and is defined by a system of 8 ordinary differential equations (ODE). The first of them is the charge conservation equation

$$C_m \dot{v} = -(I_{Na} + I_{si} + I_K + I_{K_1} + I_{K_p} + I_b) \quad (1)$$

where \(v\) is the membrane voltage measured in millivolts, \(C_m = 1 \mu F/cm^2\) is the membrane capacity. The time unit is one millisecond. The ionic transmembrane currents in the right-hand part are sodium current, slow inward current (carried by calcium ions), potassium current, inward-rectifier potassium current, plateau potassium current and background Ohmic current, measured in \(\mu A/cm^2\). They are defined by the following expressions

$$I_{Na} = G_{Na} \cdot m^3 h j \cdot (v - E_{Na})$$
$$I_{si} = G_{si} \cdot d f \cdot (v - E_{si}(v,c))$$
$$I_K = G_K \cdot x x_i(v) \cdot (v - E_K)$$
$$I_{K_1} = G_{K_1} \cdot k_1(v) \cdot (v - E_{K_1})$$
$$I_{K_p} = G_{K_p} \cdot k_p(v) \cdot (v - E_{K_p})$$
$$I_b = G_b \cdot (v - E_b) \quad (2)$$

Here \(G_q\) and \(E_q\) with \(q \in \{Na, si, K, K_1, K_p, b\}\) denote the maximal conductance and reversal potential of the corresponding ionic current. The gating variables \(g_i \in \{m, h, j, d, f, x\}\) are governed each by an ODE of the type

$$\dot{g}_i = \alpha_{g_i}(v)(1 - v) - \beta_{g_i}(v)v \quad (3)$$

The 12 nonlinear functions \(\alpha_{g_i}(v)\) and \(\beta_{g_i}(v)\) as well as \(E_{si}(v,c)\), \(x x_i(v)\), \(k_1(v)\), \(k_p(v)\) are fitted to experimental data [2]. The dynamics of the internal calcium ion concentration \(c\) is described by an ODE of the first order

$$\dot{c} = 10^{-4} I_{si}(v,d,f,c) + 0.07(10^{-4} - c) \quad (4)$$

The eight ODEs (1), (3), (4) form a closed system for the variables of state \(v, m, h, j, d, f, x, c\).

The values of the constant parameters are the same as used in [3].

This model lacks many details taken into account in more recent models, which are much more complicated.
However, it still demonstrates good qualitative and quantitative agreement with available experimental data on single-cell dynamics [2], as opposed to simplest qualitative models like the FitzHugh-Nagumo model.

2.2. Oscillatory cells and cell cultures

To describe the oscillatory activity of a cell, we modify the model by adding a constant depolarizing current to the ionic currents in Eq. (1). When its value is increased above a bifurcation value which depends upon other parameters, a limit cycle appears in the phase space of the model, thus the cell becomes oscillatory. Though this approach seems to be artificial and might not account for real physiological mechanisms of cell oscillation, the development of a more adequate model is hindered by the lack of understanding of the mentioned mechanisms in in-vitro experiments. However, in real situations, it is known that the leakage (depolarization) current of the non-pacemaker cells can increase and turn into oscillatory cells when they are dissociated from the heart tissues[4].

We model a cell culture by a square lattice with local diffusive coupling. This type of coupling represents electrical intercellular conductance coupling via gap-junctions.

The charge conservation equation (1) with the constant depolarizing current and intercellular coupling taken into account is rewritten as

\[
C_m \dot{v}_{ij} = -(I_{Na} + I_{si} + I_K + I_{K_1} + I_{Kp} + I_b) - I_{d}^0 + D \Delta_d(v_{ij})
\]

where \(i, j\) are lattice indices, \(I_{d}^0 > 0\) is the constant depolarizing current which is non-identical in different cells, \(D\) is the coupling coefficient, \(\Delta_d\) is the second-order central difference operator (discrete Laplacian).

Note, that the spatial scale of one cell in the lattice model corresponds to the characteristic scale of culture inhomogeneity rather than the size of a single cardiac cell.

3. One-dimensional case

To get some insight into the problem, we start with the one-dimensional (1D) model. We consider a chain of 400 oscillatory cells with different individual oscillation frequencies. For this we choose the current \(I_{d}^0\) in each cell randomly, uniformly distributed in the interval \([2.4; 3.2] \mu A/cm^2\). The initial conditions are chosen the same in each cell, so that the whole system initially gets depolarized simultaneously. In all our simulations we use free-end boundary conditions.

For very low coupling (for example, \(D = 0.001\)) we observe incoherent behavior with no large-scale collective activity in the culture.

When the coupling coefficient \(D\) is increased, the evolution of the system typically starts with the emergence of several pacemakers emitting waves at different frequencies, see space-time plot in Fig. 1. The subsequent evolution depends strongly on the value of the coupling coefficient.

Figure 1: Early stage of evolution in a chain of oscillatory cells \((D = 0.01)\).

At \(D = 0.01\) (and higher) only one pacemaker remains, and a globally synchronous regime sets in: all the elements are oscillating at the same period dictated by the only pacemaker, see Fig. 2 (taken after a waiting time of \(3 \cdot 10^3 ms\)).

Figure 2: Regime of global synchronization in a chain of oscillatory cells \((D = 0.01)\).

At lower values of \(D\) we observe regimes of cluster synchronization, when the chain is split into several groups - clusters, in each of them the oscillations are synchronized, but different clusters oscillate with different periods (see Fig. 3 for \(D = 0.005\)). In each cluster one can see the leading pacemaker. The greater is the value of \(D\), the larger is the size of clusters, the smaller is their number and the closer are their periods.

In Fig. 4 we plot the measured time-averaged period of oscillation versus the element number for \(D = 0.001\) (blue), \(D = 0.005\) (red) and \(D = 0.01\) (black). The
transition from asynchronous oscillation to global synchronization via cluster-synchronized regimes is clearly visible here.

4. Two-dimensional case

We carried out two-dimensional (2D) simulations in two settings. In the setting $A$ the culture consists only of oscillatory cells with different individual oscillation frequencies, same as in 1D simulations. In the setting $B$ we consider the culture as a mixture of excitable and oscillatory cells. To achieve this, we let the depolarizing currents $I_{d}$ be uniformly distributed in the interval $[0; 3.2]$. In both settings we study the character of collective activity in the culture when the coupling coefficient $D$ is varied between zero and the value when global synchronization sets in. The size of the lattice is $50 \times 50$, the initial conditions are specified the same way as in 1D simulations.

The overall picture in both settings has many features similar to that in the 1D case. In particular, for very small coupling ($D \leq 0.002$ in the setting $A, D \leq 0.01$ in the setting $B$) there is no large-scale activity in the culture. In the setting $B$ at this small coupling there are typically many independently oscillating islands separated by a non-excited background (caused by the existence of non-oscillatory excitable cells, which are not excited because of insufficient coupling strength).

For large values of coupling ($D \geq 0.005$ in the setting $A, D \geq 0.025$ in the setting $B$) the system remains with a single pacemaker emitting target waves, which propagate through the whole lattice, thus the regime of global synchronization sets in, also similar to the 1D case.

In the setting $B$ the further increase of the coupling strength $D \geq 0.1$ leads to a dying out of the oscillations. A possible intuitive explanation of this effect can be as follows. Oscillatory cells too strongly coupled to non-oscillatory excitable cells no longer can neither oscillate, nor excite the excitable cells. Thus the whole system remains at rest.

A qualitative difference from the 1D case can be observed at intermediate values of the coupling strength, for example at $D = 0.003$ (setting $A$), $D = 0.025$ (setting $B$), see sample snapshots in Figs. 5 and 6. In these cases several pacemakers are formed on the early stage, but subsequently they get destroyed, giving rise to complicated spatio-temporal regimes of activity which may exhibit spiral waves, pieces of propagating fronts and pacemakers appearing intermittently in the system. This regime persists up to very long simulation times (of the order $10^5$ ms).
5. Discussions

In the above discussions, it is clear that the coupling constant $D$ plays an important role in the global synchronization of the parameters. In real situations, this $D$ is probably the connectivity between cells. In an experiment with cardiac cell cultures from chicken embryos, Glass et al have shown that the control of connectivity of the system through the use of $\alpha$-glycerol acid and culture density can indeed produce a transition of synchronized patterns [5]. They have used a heterogeneous cellular automaton model to understand their experiment findings. It seems that heterogeneity and excitability are essential in their explanations.

In our case, the system is essentially a mixture of excitable and oscillatory cells. Although cells taken from the ventricle [1] are considered to be only excitable when they are in an intact heart, they will become oscillatory [4] after they have been dissociated from the heart tissue and plated on the culture dishes. It is known that their oscillation period will be much longer than that of the pacemakers but it is not clear what the oscillation period distribution is and whether this distribution will depend on the growth conditions. Even though all the cells might be eventually oscillatory, as a first approximation, one can still consider most of the cells as excitable as they will be driven by a few cells with the shortest oscillation periods. It is therefore reasonable to assume that there is a mixture of excitable and oscillatory cells. In this sense, we also have heterogeneity and excitability built into our systems. However, it is not clear if this heterogeneity is the same as that of Glass et al [5].

6. Acknowledgements

This research was supported by the RFBR-NSC (Project No. 05-02-90567). O.K. also acknowledges support from the “Dynasty” Foundation, Russia.

References

A Basic Theory for Available Operation of Extremely Complicated Large-Scale Network Systems

Kazuo HORIUCHI†

†Professor Emeritus, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555 Japan
Email: noriko@oishi.info.waseda.ac.jp

Abstract—In this paper, we shall describe about a basic theory based on the concept of set-valued operators, suitable for available operation of extremely complicated large-scale network systems. Fundamental conditions for availability of system behaviors of such network systems are clarified in a form of fixed point theorem for system of set-valued operators. Here, the proof of this theorem is accomplished by the concept of Hausdorff’s ball measure of non-compactness.

1. Introduction

In extremely complicated large-scale network systems, precise evaluation and perfect control, and also ideal operation, of overall system behaviors cannot be necessarily expected by using any type of commonplace technologies for maintenance, which might be accomplished by simple measure in usual hierarchical network structures.

In order to effectively evaluate, control and maintain those complicated large-scale networks, as a whole, the author has recommended to introduce some connected-block structure: i.e., whole networks might be separated into several blocks which are carefully self-evaluated, self-controlled and self-maintained by themselves, and so, which are originally self-sustained systems. However, by always carefully watching each other, whenever they observe and detect that some other block is in ill-condition by some accidents, every block can repair and sustain that ill-conditioned block, through inter-block connections, at once. This style of maintenance of system is sometimes called as locally autonomous, but the author recommends that only the ultimate responsibility on observation and regulation of whole system might be left for headquarter itself, which is organized over all blocks just as United States Government [1].

Here, let us consider complete metric linear spaces $X_i$ $(i = 1, \cdots, n)$ and $Y_j$ $(j = 1, \cdots, n)$, and their bounded convex closed subsets $X_i^{(0)}$ and $Y_j^{(0)}$, respectively, corresponding to each block, $B_i$ and $B_j$ of whole network system. Let us introduce operators $f_{ij} : X_i \to Y_j$ such that $f_{ij}(X_i^{(0)}) \subset Y_j^{(0)}$ and let $f_{ij}$ be completely continuous on $X_i^{(0)}$.

For each block $B_i$ $(i = 1, \cdots, n)$, dynamics of system behaviors can be represented originally by simple equations:

$$x_i = \alpha_i f_i(x_i), \quad (i = 1, \cdots, n), \tag{1}$$

where $\alpha_i$ is a continuous operator: $Y_i^{(0)} \to X_i^{(0)}$. These equations have solutions $x_i^*$ in every $X_i^{(0)}(i = 1, \cdots, n)$, according to the well-known Schauder’s type of fixed point theorem. Of course, these solutions represent original values of system behaviors.

But, in general conditions with inter-block connections, system dynamics may be represented in the form of system of equations:

$$x_i = \alpha_i f_i(x_i) + \sum_{j \neq i} \beta_{ij} f_j(x_j) + \sum_{j \neq i} \gamma_{ij} f_j(x_{ji}), \quad (i = 1, \cdots, n), \tag{2}$$

where $\beta_{ij}$ is a continuous operator: $Y_j \to X_i$ and $\gamma_{ji}$ is a continuous operator: $Y_i \to X_j$. In the right-hand side of this system of equations, the first term represents the original performance of the $i$-th block itself, the second term represents the operation feed-back through all other blocks ($j \neq i$) into the original $i$-th block, and the third term represents inter-block connections from all other blocks, in order to repair and sustain the $i$-th block performance. This system of equations also has solutions $x_i^*(i = 1, \cdots, n)$ in every $X_i^{(0)}$ in the $i$-th block itself, which represent resultant behaviors of block $B_i$, as a whole. (See Fig. 1)

For general situations of mutual connections between blocks, by newly introducing $n$ composition-type operators $g_i : X_i \times \Pi_i Y_j \times \Pi_i Y_j \to X_i$, where $\Pi_i Y_j$ means the direct product of $n$ $Y_j$’s for all $j \in \{1, \cdots, n\}$, and $\Pi_i Y_j$ means the direct product of $n$ $Y_j$’s for fixed $i$, we have a general system of operator equations:

$$x_i = g_i(x_i; f_{i1}(x_1), \cdots, f_{im}(x_m); f_{i1}(x_1), \cdots, f_{im}(x_n)), \quad (i = 1, \cdots, n). \tag{3}$$

Here, we can present a fixed point theorem for this general system of nonlinear operator equations [2], which is an extension of the work by Melvin [3].

However, the fluctuation imposed on the actual system is nondeterministic rather than deterministic. In this case, even the effect due to a single cause is multi-valued, and the
behavior is more naturally represented by a set of points, rather than a single point.

Therefore, it is reasonable to consider some suitable-subset of the range of system behavior, in place of single ideal point, as target which the behavior must reach under influence of system control. Now, we can name it as an “available range” or a “tolerable range” of the system behavior. Thus, by the available or tolerable range, we mean the range of behavior, in which every behavior effectively satisfies good conditions beforehand specified, as a set of ideal behaviors. From such a point of view, the theory for fluctuation imposed on the system should be developed concerning the set-valued operator.

By the set-valued operator $G$ defined on a space $X$ is meant a correspondence in which a set $G(x)$ is specified in correspondence to any point $x$ in $X$. In particular, when $G(X) \subset X$, and if there exists a point $x^* \in G(x^*)$, $x^*$ is called a fixed point of $G$.

The author has given a series of studies on set-valued operators in functional analysis aspects, and has vigorously applied it to analysis of uncertain fluctuations of network systems [4], [5], [6].

Recently, the author gave a general type of fixed point theorem for the system of set-valued operator equations, in order to treat with extremely complicated large-scale network systems [7], [8].

Namely, let us introduce $n$ set-valued operators $G_i : X_i \times \Pi^* Y_j \times \Pi^* Y_j \rightarrow F(X_i)$ (the family of all non-empty closed compact subsets of $X_i$) where $\Pi^* Y_j$ means the direct product of $n$ $Y_j$’s, for any $j \in \{1, \ldots, n\}$, and $\Pi^* Y_j$ means direct product of $n$ $Y_j$’s, for any $j \in \{1, \ldots, n\}$.

Under some natural conditions, the author presented an important fixed point theorem on the system of set-valued operator equations:

$$x_i \in G_i(x_i; f_1(x_1), \ldots, f_m(x_1); f_1(x_1), \ldots, f_m(x_m)), \quad (i = 1, \ldots, n).$$

2. A Basic Fixed Point Theorem For System of Set-Valued Operators

Now, we will present a basic theory of the fixed point theorem for such a general system of set-valued operator equations.

For the first step, let us introduce real Banach spaces $X_i \quad (i = 1, \ldots, n)$, in which the norm is represented by $\| \cdot \|$, and also their non-empty bounded closed convex subsets $X_i(0) \quad (i = 1, \ldots, n)$. Further, let us consider another real Banach spaces $Y_j \quad (j = 1, \ldots, n)$ in which the norm is represented by $\| \cdot \|$. Let us define a direct product space $V_i = \Pi^0 Y_j \times \Pi^0 Y_j$ and also let $V_i(0)$ be a non-empty bounded closed convex subset of $V_i$. Here, let us consider a vector $v_i = (x_i, \ldots, x_i; x_1, \ldots, x_n) \in V_i$ and an operator $f_i(v_i) : V_i \rightarrow Y_j$ by

$$f_i(v_i) = (f_{i1}(x_1), \ldots, f_{m1}(x_1); f_{i1}(x_1), \ldots, f_{m1}(x_m)). \quad (5)$$

Here, we know that $y_j = f_i(x_i) \in Y_j$, $y_{ji} = f_{ii}(x_i) \in Y_j$, and $y_i = (y_{1i}, \ldots, y_{ni}; y_{1i}, \ldots, y_{ni}) \in Y_i$. Therefore, we have a simple representation of the system of set-valued operators (4), as follows:

$$x_i \in G_i(x_i; f_i(v_i)), \quad (i = 1, \ldots, n). \quad (6)$$

Next, let us introduce a series of assumptions:

**Assumption 1** Let the operator $f_{ij} : X_i(0) \rightarrow f_{ij}(X_j(0)) \subset Y_j$ be completely continuous (continuous and compact).

**Assumption 2** Let the set-valued operator $G_i : X_i \times V_i \rightarrow F(X_i)$ (a family of all non-empty closed compact subset of $X_i$) satisfies the following Lipschitz condition with respect to the Hausdorff distance $d_H$: that is, there are two kinds of constants $0 < k_1 < 1$ and $h_1 > 0$ such that for any $x_{1}^{(1)}, x_{1}^{(2)} \in X_i$, for any $y_{ij}^{(1)}, y_{ij}^{(2)} \in Y_j$ and for any $y_{ji}^{(1)}, y_{ji}^{(2)} \in Y_i$, $G_i$’s satisfy inequalities:

$$d_H(G_i(x_{1}^{(1)}, y_{1j}^{(1)}, \ldots, y_{mj}^{(1)}), G_i(x_{1}^{(2)}, y_{1j}^{(2)}, \ldots, y_{mj}^{(2)})) \leq k_1 \cdot \|x_{1}^{(1)} - x_{1}^{(2)}\| + h_1 \cdot (\sum_{j=1}^{m} \|y_{ij}^{(1)} - y_{ij}^{(2)}\|) \quad (7)$$

that is,

$$d_H(G_i(x_{1}^{(1)}, y_{1j}^{(1)}), G_i(x_{1}^{(2)}, y_{1j}^{(2)})) \leq k_1 \cdot \|x_{1}^{(1)} - x_{1}^{(2)}\| + h_1 \cdot (\sum_{j=1}^{m} \|y_{ij}^{(1)} - y_{ij}^{(2)}\|) \quad (i = 1, \ldots, n), \quad (8)$$

where the norm in $Y_i$ is defined as

$$\|y_i\| = \sum_{j=1}^{m} \|y_{ij}\| + \sum_{j=1}^{m} \|y_{ji}\|, \quad (9)$$

and the Hausdorff distance $d_H$ between two sets $S_1$ and $S_2$ is defined by

$$d_H(S_1, S_2) = \max\{\sup\{d(x_1, S_2) | x_1 \in S_1\}, \sup\{d(x_2, S_1) | x_2 \in S_2\}\}, \quad (10)$$

and the Hausdorff distance $d_H$ between two sets $S_1$ and $S_2$ is defined by
where \( d(x, S) = \inf \{ ||x - y|| \mid y \in S \} \) is the distance between a point \( x \) and a set \( S \).

**Assumption 3** For any \( x_i \in X_i(0) \) and \( y_i \triangleq f_i(x_i), \) \( y_i \triangleq f_i(x_i), G_i(0)(x_i; y_{1i}, \ldots, y_{mi}; y_{1i}, \ldots, y_{mi}) \triangleq G_i(x_i; y_{1i}, \ldots, y_{mi}; y_{1i}, \ldots, y_{mi}) \cap X_i(0) \neq \phi. \)

Then, we have a final result:

**Theorem 1 (Fixed Point Theorem)** The system of set-valued operator equations
\[
x_i \in G_i(0)(x_i; f_i(v_i)), \quad (i = 1, \ldots, n)
\] has at least one fixed point \( x^*_i \in X_i(0). \)

The proof of this theorem can be accomplished by the concept of Hausdorff’s ball measure of the non-compactness, as follows:

First, we remember Hausdorff’s ball measure \( \beta(S) \) for the non-compactness of bounded subset \( S \) of real Banach space [9]:
\[
\beta(S) \triangleq \inf \{ \varepsilon \geq 0 | S \text{ can be covered with a finite number of balls of radii smaller than } \varepsilon \} \quad (12)
\]

Let \( V_i(0) \neq \phi \) be a bounded closed convex subset of \( V_i. \) Then, we can find that for any \( x_i \in X_i(0), \)
\[
\beta(G_i(0)(x_i; f_i(V_i(0)))) = 0,
\] (13)

where
\[
G_i(0)(x_i; f_i(V_i(0))) \triangleq \bigcup_{v_i \in V_i(0)} G_i(0)(x_i; f_i(v_i)).
\]

Eq. (13) means that the convex closure of \( G_i(0)(x_i; f_i(V_i(0))) \) belongs to \( \overline{F(X)} \).

Next, by Eq. (13), we can obtain a relation
\[
\beta(G_i(0)(X_i(0); f_i(V_i(0)))) \leq k \cdot \beta(X_i(0)),
\] (14)

where
\[
G_i(0)(X_i(0); f_i(V_i(0))) \triangleq \bigcup_{v_i \in V_i(0)} G_i(0)(x_i; f_i(v_i)).
\]

Now, let us introduce a sequence \( \{ W_n \} \ (n = 0, 1, 2, \ldots) \) by the successive procedure such that \( W_0 \triangleq X_i(0), \) and \( W_n \triangleq \text{conv} \ G_i(0)(W_{n-1}, f_i(V_i(0))), \ (n = 1, 2, \ldots). \) Then, from Eq. (14) we have in turn
\[
X_i(0) = W_0 \supset W_1 \supset W_2 \supset \cdots
\] (15)

and
\[
\beta(W_n) \leq k^n \cdot \beta(W_0).
\] (16)

From these relations, we see that \( \beta(W_n) \to 0, \) as \( n \to \infty. \)

Thus, for the set \( W_\infty \triangleq \bigcap_{n=0}^{\infty} W_n, \) we have
\[
G_i(0)(W_\infty, f_i(V_i(0))) \subset W_\infty,
\] (17)

which means that \( W_\infty \) is a non-empty convex compact subset invariant with respect to the operator \( G_i(0)(x_i, f_i(v_i)). \)

On the other hand, we can prove in a similar way that the set-valued operator \( G_i(0)(x_i, f_i(v_i)) \) is upper semi-continuous.

From this flow of deduction, we can see that the well known Ky Fan’s fixed point theorem on set-valued operators [10] can be applied to our problem, and, as a result, that the set-valued operator \( G_i(0)(x_i, f_i(v_i)) \) has a fixed point \( x^*_i \in X_i(0). \)

3. Concluding Remarks

The purpose of this paper is to give a mathematical foundation for available operations of complicated large-scale network systems, by means of functional-analysis-methodology, described in the forms of fixed point theorem.

The series of studies accomplished by the author, has clarified their theoretical phases, step by step, in his reports published in NOLTA2000[7], NOLTA2005[8], and to be published in NOLTA2006, in turn.

The proof of fixed point theorem in ref.[7] was accomplished by natural assumptions, on the other side, the proof of the theorem in ref.[8] was accomplished according to a refined precise deduction, in weak topology.

The third theorem, in this NOLTA report, was proved by using the concept of Hausdorff’s ball measure of non-compactness, on a more basic foundation of mathematical analysis.

**Acknowledgment**

The author wishes to express his gratitude to Professor Emeritus R. Ino, Waseda University, and Professor J. K. Hale, Brown University for their kind advice to this series of works in the mathematical aspect, at the beginning of 1980’s. This series was supported by Scientific Grants, Min. of Education, Science and Culture, Japanese Government, and also by the Special Scientific Grants, Waseda University. Finally, the author is grateful to express his gratitude to Professor S. Oishi, Waseda University, for his kind advice and support to this series of works, and further, to Dr. K. Maruyama, Visiting Lecturer, Waseda University, for the English description of manuscript including the figure, with kind advice.
References


Abstract—This paper deals with the analysis of multi-fractal signals in artificial markets generated by behaviors of multi-agents who use the Genetic Programming (GP) for learning and its applications. Agents prefer forecast models (equations) or production rules to support their decision making, and irrational agents select decisions at random. Depending on the composition of agents and the way of decisions, the signal obtained from agents’ behavior bears the multi-fractal characteristics. By applying prediction method for mono-fractal surfaces, features of the multi-fractal surfaces are extracted.

1. Introduction

There are many conventional methods for generating mono-fractals time series and surfaces, even for giving multi-fractal time series based on random numbers[1]-[3]. However, there are few methods for generating multi-fractal surfaces despite the extension of method for multi-fractal time series is partly available.

In this paper, we show the analysis of multi-fractal signals (time series and surfaces) generated by behaviors of multi-agents using the GP for learning and its applications. We emphasize the role of co-evolutionary GP and the classifier systems to emulate the real market based on the multi-agent systems[4]-[6]. In addition in the co-evolutionary learning, agents are different in owning their individual forecast model bases or just learning from public base if a agent becomes to be not satisfied with the increase of wealth. After examining the multi-fractal processes, we apply the method for feature extraction based on the prediction on mono-fractal signals proposed in previous works[10][11].

1.1. Architecture of Multi-agent System

The major component of this architecture is the heterogeneous adaptive agents. For generating multi-fractal time series, we introduce artificial stock markets. Specifically, five types of agents are defined in this architecture.

Agents of type 1 and type 3

The agents of both type 1 and type 3 are agents who forecast the value of stock price and dividend of the next period by using an adequate forecast equation model selected from a forecast model base. The difference of these two types is that the agents of type 1 possess their individual forecast model bases, but the agents of type 3 only learn from a public forecast model base, without their own bases.

Agents of type 2 and type 4

The agents of both type 2 and type 4 are agents who forecast the rise/fall of stock price and dividend of the next period by using production rules. Agents of type 4 only learn from a public forecast model base without their own bases.

Agents of type 5

Different from these agents above, the agents of type 5 seem to behave irrationally, in the way that they do not use any reasonable approaches to support their decision making process.

Learning of agents based on GP

In this paper, the cognitive behavior of adaptive agents will be modeled by applying the GP approach. Specifically, forecast equation models are evolving in responding to the market dynamics, enforced the genetic operations. In the GP, each forecast model is represented in the tree structure (called individual). In the parse tree, the non-terminal node is taken from the function sets, containing +,-,÷,×,exp, abs,sqrt

Terminal node consists of arguments chosen from set of constants.

min(t):minimum price in period[t_{time-t,t_{time-1}}]
max(t):maximum price in period[t_{time-t,t_{time-1}}]
av(t):average price in period[t_{time-t,t_{time-1}}]
price(t):stock price in period t_{time-t}
dividend(t):dividend in period t_{time-t}

We define the fitness of forecast model is the reciprocal number of squared forecast error denoted as $e_{ij}$, in which $i$ is the index of agent, $j$ is the index of forecast model in the base and $t$ accounts for time horizon.

We find many literature for the description of the GP procedure, then we omit the details[4]-[6][8][9].
2. Fractality of Stock Price

For testing the time series to be fractal, we utilize the logarithm of the wavelet transform coefficients of stock price by depicting the value along the axis of dilation index $D$. Namely, the spectrum of the time series is described by using the variance $\sigma^2$ of the time series, and the fractal dimension $D$.

\[
S(\omega) = \sigma^2 \omega^{-\gamma},
\]

\[
\gamma = 5 - 2D.
\]

Now, we assume that the time series is represented by using the orthogonal Wavelet transform.

\[
x(t) = \sum_{n} \sum_{m} x_{mn}^m \psi_{mn}^m(t),
\]

\[
x_{mn}^m = \int_{-\infty}^{\infty} x(t) \psi_{mn}^m(t) dt.
\]

where $\psi_{mn}^m(t)$ is defined by using the scale and shift transform of the basic function $\psi(t)$. The numbers $n, m$ mean the indices of the scale and shift transform (dilation and transform), respectively. Since $x(t)$ has the fractal geometry, we have next relation for coefficient $x_{mn}^m$ based upon the characteristics of the Wavelet function.

\[
\text{var}(x_{mn}^m) = \sigma^2 2^{-2\gamma m}.
\]

By taking the logarithm of equation (5), we have a linear function of index $m$. Then we adopt a linear regression curve to the logarithm of variance, and define the root mean square error of the difference as

\[
R_s = \left[ \sum_m (\log(var(x_{mn}^m)) - c_0 - c_1 m)^2 \right]^{1/2} / (MX.),
\]

where $M$ is the range of index $m$, and $X_r$ is the difference between the maximum and minimum of $\log(var(x_{mn}^m))$. If the time series $x(t)$ is fractal, then $R_s=0$ theoretically.

Composition of agents

We consider six cases of composition of agents where $N_i$ mean the number of agents of type $i$.

Case 1: $N_i=100,100,100,100,100$

Case 2: $N_i=100,100,0,100,100$

Case 3: $N_i=0,100,100,100,100$

Case 4: $N_i=100,100,0,0,0$

Case 5: $N_i=0,0,100,100,100$

Case 6: $N_i=0,0,0,100,100$

Then, $R_s$ for Case $i, i = 1 \sim 6$ are 0.0064, 0.0081, 0.0041, 0.0040, 0.053, 0.0069, respectively. We see that the artificial stock prices deviate from mono-fractal processes in each cases except for case 6.

Role of LCS

For the test of fractality, the randomness in agents’ behavior is expected to give good effect on the reality of artificial stock price. We employ a method for the explanation of LCS (Learning Classifier Systems) effect on the stock price by changing the range of selection of individuals from the pool[7].

We define $p_S$ as a probability in selecting an individual from the pool. We assume agents of type 1, 2, 3, 4 use one of the individual at random from $N_S$ individuals which are selected from the top of $N_S$ members in the pool. The probability $p_S$ is therefore defined as $p_S = N_S / N_G$. If $p_S = 1$ the method for the usage of individual is the original LCS scheme. If $N_S = 1, p_S = 1/N_G$, then the method corresponds to the exclusion of LCS scheme, in which agents use only individuals with highest fitness.

The result for several $p_S$ ($p_S = 1.0, 0.8, 0.6, 0.4, 0.2, 0.0$ ) is summarized as $R_s=0.0065, 0.0078, 0.091, 0.014, 0.024, 0.014$, respectively.

As is seen from the result, if $p_S < 0.6$, the artificial stock price is no more mono-fractal.

3. Characterizing multi-fractal by WTMM

In the following discussion, we use the Wavelet Transform Modulus Maxima (WTMM) method proposed by Muzy and Bacry for finding the multi-fractal processes in time series [3]. At first, we calculate the wavelet transform of the time series $x(t)$.

\[
z(b,a) = \frac{1}{a} \int_{-\infty}^{\infty} \phi \left( \frac{x-b}{a} \right) x(t) dt.
\]

where $\phi(t-a)/a$ is obtained from the basic wavelet function $\phi(t)$ by using the translation index (scale parameter) $a$ and the dilation index (space parameter) $b$. Then, the multi-fractal processes can be characterized by using the maxima of the $q$-th moment of the $z(b,a)$. We define following function.

\[
Z(a,q) = \sum_{l[L(i)]) \sup_{[y,y_0]} z(y,a) \right]^{q/2}.
\]

where $l$ denotes the lines called as the maxima lines, and are obtained as a set of points by assigning $a$ to a fixed value and varying $b$ as variable $y$ giving the maximum value of $Z(a,b)$. Then, $L(a)$ denotes the set of these lines. After these calculation of maxima, we obtain the exponent $\tau(q)$ in the limit $a \rightarrow 0$ as follows.

\[
Z(a,q) \rightarrow a^{\tau(q)}.
\]

Then, we can compute the singularity spectrum $D(h)$ from the Legendre transform of $\tau(h)$.

\[
D(h) = \min_{q} \{qh - \tau(q)\}.
\]

The singularity spectrum $D(h)$ of the signal is then defined as the function that gives, for a fixed $h$, the Hausdorff dimension of the set of points of $x$ where the exponent $H(x)$ is equal to $h$. If the curve $D(h)$ spreads along $h$, then the time series bears multi-fractality, otherwise the curve converges to a single point.

Fig.1 shows an example of $D(h)$ for Case 1 and Case 2, and we see multi-fractility for each case.

136
4. Generation of multi-fractal surfaces

Then, we extend the method for generating fractal time series to then generation of fractal surfaces. In previous discussion, we assume that agents invest and trade only one kind of stock. Then, we assume following markets consisting multiple stocks (issues).

(1) sorting issues according to same industries

It is assumed that stocks issues are sorted into a series (list L) where resemble issues are placed in their neighborhood. For example, stocks that are issued by companies belonging to the same industry are placed on the neighborhood in the list L. We also assume that the end of list L is combined with another end of list L like a torus.

(2) agents invest to multiple issues

We assume that agents are allowed to invest to multiple issues according to list L. An agent mainly invest to i th issue on list L, but simultaneously, the agent invest to i + is, i - is th issues on list L. If i + is exceeds the length of L or i - is is less than 0, then agents switch them along the torus.

(3) decision of buy/sale of issue

Agents predict future price of issues based on the previous prices of issues. Then, the arguments such as min(t), max(t) for a single issue are extended to min(t, j), max(t, j) for the prediction of future price of j th issue. These extended arguments are used in the prefix representation of individuals for GP learning. Based on the prediction of future prices of issues, agents decide to buy/sale of issues and the volumes to be traded.

(4) fractal surface

Then, we summarize the data for i th issue at time t as y(t, i) after sufficient time of learning by agents. As a result, we find that the two-dimensional data y(t, i) reveals as a fractal surface. After precise examination, we also find that x(t, i) exhibits the characteristics of multifractality.

Fig.2 shows an example of x(t, i). Fig.3 shows an example of log[var(x^n)] (the logarithm of time series x(t) = y(t, i) for a certain fixed i). As is seen from Fig.2, the segment of surface x(t, i) bears fractality, and then it is expected that x(t, i) becomes an fractal surface.

5. Feature extraction using Prediction

Since the multi-fractality is characterized by a short-term feature rather than the long-term dependency of the mono-fractal signals. Then, we employ the smoothing method of the mono-fractal time series for detecting the feature of multi-fractality.

The smoothing method has been proposed in previous works, and the overview is summarized as follows[10][11]. Suppose time series s(x(t) is represented as

\[ x(t) = \int_{t_0}^{t} h(t, \tau) x(\tau) d\tau, \quad t > t_0. \]  \hspace{1cm} (11)

where the impulse response function h(t, \tau) is represented by the scaling function \( \phi(t) \)

\[ h(t, \tau) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} h_i \phi_{N_i}(t) \phi_{N_j}(\tau). \]  \hspace{1cm} (12)
Then, we can apply the same expansion for the surface as
\[
\tilde{z}(x, y) = b^{-2} \int_0^{\phi_{N_i}(t)} \int_0^{\phi_{N_j}(t)} h(x - \tau, y - \mu) z(\tau, \mu) d\tau d\mu.
\]
\[
b = a^D, a = T_2/T_1, T_2 > T_1.
\]
(19)

We obtain distinctive feature points for artificial stock prices by using the method. Figure 4 shows an example for detecting feature points.

Fig.5 shows an example of original artificial stock price and mono-fractal prediction which is giving us the feature extraction of fractal time series.

References


Robust Stochastic Direct Adaptive Radial Basis Neural Networks Synchronization of Chua’s Circuits with Unknown Function

I. Saboori, M.B. Menhaj, K. Nikravesh and B. Karimi

Abstract—This paper presents a robust stochastic direct adaptive radial basis neural network synchronization of Chua’s circuit. Based on Lyapunov stability theory, an adaptive radial basis neural network is tuned and used in a nonlinear robust controller to make the states of two identical Chua’s systems with a bounded stochastic noise and fully unknown Chua’s diode characteristic nonlinear function, synchronized. It is worth mentioning that only one of the three states of master system is observed and used in control laws. Finally, some numerical simulations are performed to illustrate the efficiency and robustness of the proposed robust adaptive synchronization scheme.

I. INTRODUCTION

Since its introduction by Pecora and Carrol in 1990 [1], chaos synchronization has received increasing attention due to its theoretical challenge and its great potential applications in secure communication, chemical reaction, biological systems and so on [2]. Basically, chaos synchronization problem can be formulated as follows. Given a chaotic system, which is considered as the master (or driving) system, and another identical system, which is considered as the slave (or response) system, the aim is to force the response of the slave system to synchronize the master system [3]. Great efforts have been devoted to achieving this goal in the last few years and a large variety of approaches has been proposed, such as sampled-data feedback synchronization method [4], impulsive control method [5], adaptive design method [6], and invariant manifold method [7].

In this paper, a new direct adaptive controller is developed for Chua’s system using radial basis neural networks and Lyapunov stability technique. In this work the nonlinearity term of the master system dynamics is fully unknown. It is shown that the tracing error between master and slave systems converges to zero in presents of a bounded stochastic noise in master system dynamics.

The paper is organized as follows. Section 2 presents the problem statement. A few literatures on the radial basis neural networks and linear parameter approximators are described in section 3. The procedure of control design is described in section 4 and global stability of the overall system is proved by the theorem that established in this section. This section followed by simulation results in section 5 to show the efficiency of presented method. Finally, section 6 concludes the paper.

II. SYSTEM DESCRIPTION

Consider the chaotic system of Chua’s oscillator given by the below equations [8]:

\[
\begin{align*}
\dot{x} &= -px + py - pg(x) + \Omega(x, y, z, t) \\
\dot{y} &= x - y + z \\
\dot{z} &= -qy - rz 
\end{align*}
\]

Where \(x, y, z\) are states, \(p, q, r > 0\) are the system parameters, \(g(x)\) is a unknown nonlinear function which describe the voltage-current characteristics of the nonlinear resistor called Chua’s diode and \(\Omega(x, y, z, t)\) represents the stochastic noise and deterministic disturbance. We suppose that the \(\Omega(x, y, z, t) \in \mathcal{H}_p\) is bounded by

\[
\|\Omega(x, y, z, t)\|_p \leq \rho 
\]

In order to observe the direct adaptive neural network synchronization behavior in the Chua’s system, we assume that we have two Chua’s systems so that the master system governed by equation (1) has uncertainties and an unknown nonlinear function, and a second system (the slave system) is described by

\[
\begin{align*}
\dot{x} &= -px + py + u \\
\dot{y} &= \hat{x} - y + z \\
\dot{z} &= -q\hat{y} - rz 
\end{align*}
\]

Where \(\hat{x}, \hat{y}, \hat{z}\) the states of slave system, and \(u\) represents the adaptive neural control input, which will be defined later for the purpose of synchronization the two Chua’s systems (1),(3).

We define the synchronization errors between the states of master system and the slave systems as:

\[
\begin{align*}
e_x &= \hat{x} - x \\
e_y &= \hat{y} - y \\
e_z &= \hat{z} - z
\end{align*}
\]

So, we can easily drive the synchronization error dynamical system from (1),(3) and (4):

Manuscript received April 11, 2006.

Iman Saboori is MSc student of the Amirkabir Technology, Tehran, Iran (e-mail: iman_saboori@aut.ac.ir).

M.B. Menhaj is full professor of the Amirkabir Technology, Tehran, Iran (e-mail: menhaj@aut.ac.ir).

K. Nikravesh is full professor of the Amirkabir Technology, Tehran, Iran (e-mail: nikravesh@aut.ac.ir).

Bahram Karimi is PhD student the Amirkabir Technology, Tehran, Iran (e-mail: bahram-karimi@aut.ac.ir).
\[
\dot{e}_e = -pe_e + pe_y + f(x) + u - \Omega(x,y,z,t) \\
\dot{e}_y = e_x - e_y + e_z \\
\dot{e}_z = -q e_x - re_z
\]
Where \( f(x) \triangleq pg(x) \). We can rewrite the synchronization error dynamical system as follow:
\[
\dot{e} = A e + b(f(x) + u - \Omega(x,y,z,t))
\]
Where \( e \triangleq [e_x, e_y, e_z]^T \) and the matrix \( A \) and the vector \( b \) are:
\[
A \triangleq \begin{bmatrix}
-p & p & 0 \\
1 & -1 & 1 \\
0 & -q & -r
\end{bmatrix}, \quad b \triangleq \begin{bmatrix}
1 \\
0 \\
0
\end{bmatrix}
\]

### III. RADIAL BASIS NEURAL NETWORKS

A radial basis neural network (RBNN) is made up of a collection of parallel processing units called nodes. The output of the \( i \)th node is defined by a Gaussian function as:
\[
z_i(x) = e^{-\frac{(x - c_i)^T (x - c_i)}{\sigma_i^2}}
\]
where \( x \) is the input to the network, \( c_i \) is the center of the \( i \)th node, and \( \sigma_i \) is its size of influence. The output of a radial basis network, \( y = F(x,\psi) \), may simply be calculated by either a summation so that
\[
F(x,\psi) = \sum_{i=1}^{n} \psi_i z_i(x)
\]
or by a weighted average
\[
F(x,\psi) = \frac{\sum_{i=1}^{n} \psi_i z_i(x)}{\sum_{i=1}^{n} z_i(x)}
\]
Where \( \psi = [\psi_1,\psi_2,\ldots,\psi_n]^T \) is a vector of network weights. We notice that (9) and (10) may be rewritten as
\[
\tilde{\zeta}(x) = [z_1(x),\ldots,z_n(x)]^T
\]
from (11) for the weighted average (9), and; defined by (12) for the weighted average (10).
\[
\tilde{\zeta}(x) = [z_1(x),\ldots,z_n(x)]^T / \sum_{i=1}^{n} z_i(x)
\]

Given a single RBNN, it is possible to approximate a wide variety of functions simply by making different choices for \( A \). In particular, if there are a sufficient number of nodes within the network, then there exists some \( \psi^* \) such that
\[
\sup_{x \in S_x} | F(x,\psi^*) - f(x) | \leq \varepsilon_0
\]
Where \( S_x \) is a compact set, and \( \varepsilon_0 > 0 \) is a finite constant provided \( f(x) \) is continuous [9]. This lets us express \( f(x) = F(x,\psi^*) + e(x) \) with \( e(x) \leq \varepsilon_0 \) when \( x \in S_x \).

Notice that even though RBNN’s are linear in a set of adjustable parameters, we may, e.g., approximate a function \( f(x) = a + \cos(bx^T x) \) which is not linear in an independent set of parameters \( [a b]^T \). Thus we are using an approximator which is linear in the parameters to describe functions which are not necessarily linear in another set of parameters. Even though we will be defining the control laws in terms of radial basis networks, it should be noted that any universal approximator which is linear in the adjustable parameters may be considered. Other examples are standard fuzzy systems with adjustable output membership centers [10], Takagi–Sugeno fuzzy systems [11], CMAC networks, among others.

### IV. DIRECT ADAPTIVE NEURAL NETWORK SYNCHRONIZATION

The controllability matrix for \( A, b \) is:
\[
\begin{bmatrix}
b & Ab \\
A b & A^2 b
\end{bmatrix} = \begin{bmatrix}
1 & -p & p^2 + p \\
0 & 1 & -p + 1 \\
0 & 0 & -q
\end{bmatrix}
\]

Because \( q \neq 0 \) the controllability matrix is full rank and we can find the vector \( K \) such that the following matrix is Hurwitz:
\[
\Lambda = A - bK^T < 0
\]
But, we only can observe \( x \) from the state variables of the master system. So, we use \( K = [a \ 0 \ 0]^T \). From (16), it is easy to see that the \( \Lambda \) is Hurwitz.
\[
\Lambda = A - bK^T = \begin{bmatrix}
-(p + a) & p & 0 \\
1 & -1 & 1 \\
0 & 0 & -q
\end{bmatrix}
\]

By defining \( u \) as
\[
u = -aw_x + u_x
\]
Where \( u_x \) is the neutral control term. We can obtain (18) from (6), (15):
\[
\dot{e} = \Lambda e + b(u_x + f(x) - \Omega(x,t))
\]
Because the \( f(x) \) is an unknown function, we use a radial basis neural network \( F(x,\psi) \) with the following structure to approximate it.
\[
F(x,\psi) = \psi^T \tilde{\zeta}(x)
\]
We define the optimum network weights vector \( \psi^* \) as:
\[
\psi^* = \arg \min_{\psi} \sup_{x} | F(x,\psi) - f(x) |
\]
So, we can write the following equation:
\[
f(x) = F(x,\psi^*) + e(x)
\]
Where \( e(x) \) is the representation error and depends on the network structure. Assume that representation error \( e(x) \) is bounded by \( \varepsilon_0 > 0 \), i.e., \( |e(x)| \leq \varepsilon_0 \).
An adaptive algorithm will be defined to estimate $\psi^*$ with $\psi$. This estimate is then used to define the control law $u_n$ as:

$$u_n = -\psi^T \xi(x) - 2\rho_2 \frac{e_\xi}{|e_\xi|}$$

(22)

Where $\rho_2 = \rho + e_\xi$.

**Theorem:** The synchronization error dynamical system (6) will be asymptotically stable using the control law (23) and network parameter updating rule (24).

$$\psi = \Gamma e_\xi \xi(x)$$

(24)

Where $\Gamma$ is a symmetric positive definite matrix.

**Proof:** From the error dynamical system (6) and (23) we can drive (25):

$$\dot{\xi}(x) = \Lambda - b\psi^\top \xi(x) + b\psi + bw(e) - b\Omega(x,t)$$

(25)

Where $\psi = \psi - \psi^*$ is the network parameter estimation error and $w(e)$ is

$$w(e) = -2\rho_2 \frac{e_\xi}{|e_\xi|} = -2\rho_2 \frac{e^\top \psi P \psi}{|e_\xi|}$$

(26)

In (24) $P$ is a symmetric positive definite matrix, which defined in (27) and satisfies the Lyapunov equation (28) for the following positive definite matrix $Q$.

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & p & 0 \\ 0 & 0 & p/q \end{bmatrix}, \quad Q = \begin{bmatrix} 2(p + a) & -2p & 0 \\ -2p & 2p & 0 \\ 0 & 0 & 2rp/q \end{bmatrix}$$

(27)

$$PA + A^\top P = -Q$$

(28)

Consider the positive Lyapunov function $V(e, \psi, t)$ as

$$V(e, \psi, t) = e^\top P e + \psi^\top \Gamma^{-1} \psi$$

(29)

The time derivative of $V$ along with the synchronization error trajectory (25) is

$$\dot{V}(e, \psi, t) = e^\top (PA + A^\top P)e - 2\psi^\top P \psi \xi(x)$$

$$+ 2e^\top P \psi w(e) - 2e^\top P \psi d(x, t) + 2\psi^\top \Gamma^{-1} \dot{\psi}$$

(30)

Where $d(x, t) = -e(x) + \Omega(x, t)$. Using the fact that $\psi = \psi^*$ and (28), we can obtain (31) from (30).

$$\dot{V}(e, \psi, t) = -e^\top Q e + 2\psi^\top (-e^\top P \psi \xi(x) - \Gamma^{-1} \dot{\psi})$$

$$+ 2e^\top P \psi w(e) - 2e^\top P \psi d(x, t)$$

(31)

Knowing that $|d(x, t)| \leq |e(x)| + |\Omega(x, t)| \leq \epsilon_0 + \rho = \rho_2$ and using the network parameters adaptive rule (24) and equation (26) we can drive:

$$\dot{V}(e, \psi, t) \leq -e^\top Q e$$

(32)

V. SIMULATION

Consider the dimensionless version of the Chua’s master system (1) with unknown nonlinear function $g(x)$ and the slave Chua system (3). We use the following function as $g(x)$ to perform the simulation results. In particular, it is worthy to noting that the only measurable state is $x(t)$.

$$g(x) = bx + 0.5(a - b)(|x + 1| - |x - 1|)$$

(33)

Where. $a = -8/7, b = -5/7$. The following set of parameters that exhibits chaotic behavior of system (1) is employed

$$p = 10, \quad q = 0.0385, \quad r = 16$$

(34)

Meanwhile, the initial conditions of the master and slave systems are taken as

$$x = [0 \ 0 \ \ 0.1]^T, \quad \dot{x} = [7 \ -9 \ \ 5]^T$$

(35)

The noise signal is of the form

$$\Omega(x, y, z, t) = \frac{1}{2}\sin(2t) + \cos(x + y + z)$$

(36)

Figure 1 shows the chaotic attractor of the master system. Figure 2 shows the state evolution of the master and slave systems. Figure 3 depicts the tracking errors and clearly shows that the synchronization errors are converging to zero quickly, and Figure 4 illustrates the bounded ness of the radial basis neural network parameters. In the following simulation the controller parameters are as follows.

$$\rho_2 = 1.6, \quad \Gamma = 5000I, \quad n = 20$$

(37)

Where $n$ is the nodes number of neural network.
VI. CONCLUSION

This paper has addressed a robust adaptive RBFNN controller design for the synchronization problem of Chua’s systems with unknown diode nonlinear function and bounded noise. It is worth mentioning that only one of the three states of master system is observed and used in control laws. The proposed robust adaptive controller scheme has also been successfully applied to synchronize asymptotically both master and slave chaotic systems. The numerical simulation indicates the effectiveness of the proposed synchronization scheme.

REFERENCES

INSTABILITIES OF SYNCHRONIZATION IN ENSEMBLES OF CIRCLE MAPS

G.V. Osipov, M.V. Ivanchenko† and J. Kurths‡

†Radiophysical Dept., University of Nizhny Novgorod Gagarin ave. 23, Nizhny Novgorod, Russia, 603950
‡Institute of Physics, University Potsdam, 10, Am Neuen Palais, D-14415, Potsdam, Germany
Email: osipov@rf.unn.ru, ivanchenko@rf.unn.ru, juergen@agnld.uni-potsdam.de

Abstract— Stability conditions of synchronous regime in the chains of coupled circle maps are studied. It was shown that with increasing coupling (i) the transition from a non-synchronous state to a synchronous one occurs through the long-wave bifurcation and (ii) the transition from a synchronous state to a non-synchronous one occurs through the short-wave instabilities. Influence of asymmetry of coupling was analyzed.

1. Introduction

In this paper we study conditions for an onset and breakup of the chaotic synchronization regimes in ensembles of coupled circle maps (CMs). For networks of coupled maps different problems of synchronization, pattern formation, and spatio-temporal chaos have been investigated [1-6]. Synchronization in ensembles of coupled CMs has found important practical applications in electronics, radio-engineering or communications, in particular, in networks of digital phase-locked loops [7-12].

2. Common model

The main model is a chain of identical CMs which are locally diffusively coupled [12]:

\[
\phi_{j+1}^{k} = \omega + \phi_{j}^{k} - F(\phi_{j}^{k}) + d_{1}(\sin(\phi_{j+1}^{k} - \phi_{j}^{k})) + d_{2}(\sin(\phi_{j-1}^{k} - \phi_{j}^{k})).
\]  

(1)

Here \(j = 1, ..., N\) corresponds to the number of individual CMs and \(d_{1}, d_{2}\) are the coupling strengths. The parameter \(\omega\) characterizes the individual frequency of rotation. The function \(F(\phi_{j}^{k})\) is a piecewise linear \(2\pi\)-periodic function of the form

\[F(\phi_{j}^{k}) = c\phi_{j}^{k}/\pi\]

(2)

defined in the interval \((-\pi, \pi]\), and \(c\) is the control parameter. We consider negative values \(c\). In this case the eigenvalue of the autonomous map \(\rho_{0} = 1 - F(\phi_{j}^{k}) = 1 - c/\pi > 1\) and therefore the rotations are chaotic.

We assume that the system is subjected to (i) periodic boundary conditions:

\[\phi_{j}^{k} = \phi_{j+N}^{k}\]

(3)

(ii) or free-end boundary conditions:

\[\phi_{0}^{k} = \phi_{1}^{k}; \quad \phi_{N+1}^{k} = \phi_{N}^{k}.
\]

(4)

3. Symmetrically coupled maps

First we study the case of symmetrically coupled maps, i.e. \(d_{1} = d_{2} = d\). In homogeneous chains of chaotic oscillators the regime of complete chaotic synchronization is possible. This regime is realized if the state

\[\phi_{1}^{k} = \phi_{2}^{k} = ... = \phi_{N}^{k} = \bar{\phi}^{k}\]

(5)

is stable.

The specific type of the considered function \(F(\phi)\) (2) allows to find analytically the stability conditions of the synchronous state.

Linearizing around the uniform solution (5), the system (1) for deviations \(\xi_{j}^{k} = \phi_{j}^{k} - \bar{\phi}^{k}\):

\[\xi_{j}^{k+1} = \xi_{j}^{k} + F'(\bar{\phi}^{k})(1 - 2d)(\xi_{j}^{k} + d(\xi_{j-1}^{k} + \xi_{j+1}^{k})\lambda)
\]

(6)

\(j = 1, ..., N\).

or in the vector form:

\[\xi^{k+1} = J\xi^{k},\]

(7)

where \(\xi^{k} = [\xi_{1}^{k}, \xi_{2}^{k}, ..., \xi_{N}^{k}]^{T}\), and the Jacobian matrix \(J\) is

\[
J = \begin{pmatrix}
\rho_{0}\bar{\phi} & d & 0 & \cdots & \delta d \\
-1 & \rho_{0}\bar{\phi} & d & \cdots & 0 \\
0 & -1 & \rho_{0}\bar{\phi} & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
\delta d & 0 & \cdots & \cdots & \rho_{0}\bar{\phi}
\end{pmatrix}
\]

where \(\bar{\phi} = 1 - 2\bar{\phi}\), and \(\delta = 1\) for periodic boundary conditions and \(\delta = 0\) for free-end boundary conditions.

Because \(\rho_{0}\) is constant, the matrix \(J\) consists of constant coefficients, i.e. it does not depend on \(k\). This allows us to find the eigenvalues of the solution (5). One of them is always equal to the eigenvalue of the individual map \(\rho_{0}\). We remind that \(\rho_{0} > 1\) as \(c < 0\). While \(\rho_{0}\) manifests chaotic rotations, the stability of the synchronized state is determined by the other \(N - 1\) eigenvalues. If \(|\bar{\phi}| < 1\) for all \(I = 1, 2, ..., N - 1\), then state (5) is stable.

Let us consider separately both boundary conditions.
### 3.1. Periodic boundary conditions

For periodic boundary conditions the eigenvalues read:

$$\rho_I = \rho_0 - 2d(1 - \cos \psi_I)$$

$$I = 0.1, \ldots, N - 1,$$

where $\psi_I = 2\pi I/N$. If $d = 0$, i.e. the maps are uncoupled, hence all $N$ eigenvalues are equal to $\rho_0$. With increasing $d$, all eigenvalues except $\rho_0$ decrease and enter the interval $(-1; 1)$. The largest among them are $\rho_1 = \rho_{N-1}$. The smallest eigenvalues are $\rho_0$ for even $N$ and $\rho_{(N-1)/2}$ for odd $N$. Therefore in order to obtain the stability region of the synchronous state, it is sufficient to look on the eigenvalues $\rho$ and $\rho_0/2$ for even $N$, and $\rho_1$ and $\rho_{(N+1)/2}$ for odd $N$. Then the synchronization region in (1) can be determined from the inequalities:

$$\rho_1 < \rho < \rho_0/2$$

for even $N$, and

$$\rho_1 < \rho < \rho_0 + \rho_0/2$$

for odd $N$, (9)

which provides the interval of the coupling parameter $d$ that corresponds to the synchronous regime:

(i) for even $N$

$$d^* \equiv \frac{c}{2\pi(1 - \cos(2\pi/N))} < d < \frac{1}{2} - \frac{1}{4\pi} \equiv d^*$$

(ii) for odd $N$

$$d^{**} \equiv \frac{c}{2\pi(1 - \cos(2\pi/N))} < d < \frac{2c}{4\pi(1 - \cos((N-1)/N))} \equiv d^*:$$

These conditions show that the stability of the synchronous state is determined by (a) eigenvalue of the single map, and (b) by the number of coupled maps. In Fig. 1 the dependencies of $d^{**}$ and $d^*$ on the parameter $c$ for different $N$ is shown. One can see that with increasing $N$ the synchronization region $S_p$ becomes smaller.

From (10) for a chain of fixed length $N$ one can calculate the maximal value of $c^{**}$, such that at $c \leq c^{**}$ a synchronous regime is impossible for any coupling. For even $N$ it is (Fig. 2):

$$c^{**} = \frac{\cos(2\pi/N) - 1}{\cos(2\pi/N) + 1}$$

With increasing coupling at the transition to a synchronous regime the last unstable eigenmode corresponds to the $l = 1$ eigenvalues $\rho_1$. In this case $\psi_1 = 2\pi/N$. The long-wave bifurcation takes place. For $d$ a bit less than $d^{**}$ the distribution of $\phi_k^h$ for fixed $k$ will be sinusoidal: $a\sin(2\pi j/N) + b$, where $a$ and $b$ depend on $k$.

For even $N$ the first mode which becomes unstable is $l = N/2$. Therefore for $d$ a bit larger than $d^{**}$ the ”antiphase” ($\psi_{N/2} = \pi$) distribution, at which the values $\phi_k^h$ for all $k$ are identical for all even elements, i.e. $\phi_{2j}^h = \phi_{even}$ as well as for all odd elements, i.e. $\phi_{2j+1}^h = \phi_{odd}^h$. The short-wave bifurcation takes place.

These and other distributions $\phi_j$ in the chain (1) with periodic boundary conditions at $N = 256, c = -0.00001$ for different couplings $d$ after 10000 iterations are shown in Fig. 3. For chosen parameters with increasing $d$ the transition ”non-synchronous regime - synchronous regime” occurs at $d^{**} \approx 0.2113$. The transition ”synchronous regime - non-synchronous regime” takes place at $d^* \approx 0.5$. Sinusoidal and ”antiphase” distributions of $\phi_j$ are presented in Fig. 3(a) and in Fig. 3(b), correspondingly. With further increasing $d$ a modulation of $\phi_j$ takes place. (Fig. 3(c)). Then more and more eigenvalues become smaller than $-1$, i.e. more and more corresponding modes become unstable, which leads to more complex distribution of $\phi_j$ (Fig. 3(d, e)).

This analysis allows to identify only local stability of synchronous states. Due to the homogeneity of the chain

![Figure 1](image1)

Synchronization regions in the chain (1) with periodic boundary conditions (region $S_p$ bounded by dashed lines) and free-end boundary conditions (region $S_f$ bounded by solid lines) for $N = 10(a), N = 20(b), N = 50 (c)$.

![Figure 2](image2)

Figure 2: Dependence of the critical parameter value $c^{**}$ on the number of elements $N$ in the chain (1) with periodic and free-end boundary conditions. Synchronization takes place in regions under the curves.
other regimes are possible too, i.e. the system (1) is multistable. So there exist stable splay-state distribution with linear distribution $\phi_j = a \sin(2\pi j/N) + b$. $d = 0.07$ (dashed line corresponds to homogeneous synchronous state) (a), $d = 0.508$ (“antiphase” distribution) (b), $d = 0.748$ (c) (values $\phi_j$ only for even $j$ are shown) , $d = 0.88$ (d), $d = 0.91$ (e).

3.2. Free-end boundary conditions

Now we study free-end boundary conditions: $\xi_0 = \xi_1$ and $\xi_{N-1} = \xi_N$ in (6). The eigenvalues in this case are:

$$\rho_I = \rho_0 - 2d(1 - \cos(\psi_I))$$
$$I = 0, 1, ..., N - 1,$$ (13)

where $\psi_I = \pi I/N$. As for periodic boundary conditions the largest eigenvalue after $\rho_0$ is $\rho_1$, but the smallest one is now $\rho_{N-1}$. Again one can calculate the stability region $S_f$ of the synchronous state:

$$d^* = \frac{c}{2 \sin(\pi/N)} < d < \frac{2c - c}{2 \sin(\pi/N)}$$ (14)

Such regions $S_f$ for different $N$ are presented in Fig. 1 which shows that for chains of fixed length the stability region of the synchronous state in chains with periodic boundary conditions $S_p$ is essentially larger compared to those with free-end boundary conditions $S_f$.

The critical value $c^{cr}$ is:

$$c^{cr} = \pi(1 - \frac{1}{\cos(\pi/N)})$$ (15)

This curve lies below the curve $c^{cr}$ (12) for periodic boundary conditions (Fig. 2), i.e. we find again that for chains closed in a ring the synchronous regime can exist for larger size (larger $N$) than in chains with free ends.

In Fig. 5 the distributions of $\phi_j$ in the chain (1) with free-end boundary conditions for $N = 256, c = -0.0001$ for different $d$ after 10 000 000 iterations are shown. At the transition to a synchronous regime at $d$ a bit less than $d^*$ the last unstable eigenmode corresponding to $I = 1$ is sinusoidal $a + b \sin(\pi j/N)$ (where $a, b$ depend on $k$) (Fig. 5(b)). In long chains at the transition to a non-synchronous regime at $d$ a bit larger than $d^*$ the first stable eigenmode corresponding to $I = N - 1$ is close to the “antiphase” ($\phi_{N-1} \approx \pi$ (Fig. 5(c)). With increasing $d > d^*$ the distributions of $\phi_j$ becomes more complex.

Therefore for both boundary conditions the transition from a synchronous state to non-synchronous one occurs through long- and short-wave instabilities [13].

4. Effect of asymmetry of coupling

In the case of asymmetrically coupled maps (Eq.(1)) the stability conditions can be also found analytically. In Figs.
Figure 5: Distributions $\phi_j$ in the chain (1) with free ends for $N = 256$, $c = -0.0001$ for different coupling strength $d$ after 10 000 000 iterations. $d = 0.01$ (a), $d = 0.211$ (b) (distribution $\phi_j = a \sin(\pi j/N) + b$, $d = 0.501$, (c), $d = 0.864$ (d), $d = 0.99$ (e).

6-7 the synchronization regions for both boundary conditions are presented. One can see that for periodic boundary conditions the asymmetry of coupling leads to decrease of synchronization region, but for free-end boundary conditions the asymmetry is a reason of the enhancement of synchronization.

Figure 6: Synchronization regions $S$ in the chain (1) of asymmetrically coupled CMs with periodic boundary conditions for $c = -0.07$ (a), $c = -0.17$ (b), and $c = -0.5$ (c). $N = 10$.

5. Acknowledgements

This work was supported by RFBR (Project No. 06-02-16596), RFBR-NSC (Project No. 05-02-90567) and RFBR-MF (Project No. 05-02-19815). M.I. acknowledges support of the "Dynasty" foundation.

References

A New Spatiotemporal Chaotic System with Advantageous Synchronization and Unpredictability Features.

Soumaya Meherzi°*, Sylvie Marcos* and Safya Belghith°

°SYSCOM/ENIT, BP 37 Tunis Belvédère 1002, Tunisia.
*LSS/SUPELEC/CNRS, Plateau de Moulon, 91192 Gif-sur-Yvette, Cedex, France.
Email: soumaya.mehrzi@lss.supelec.fr, sylvie.marcos@lss.supelec.fr, and safya.belghith@enit.rnu.tn

Abstract– We address spatiotemporal chaotic systems as symbol sequences generators. We propose a new family of spatiotemporal systems which are advantageous in terms of synchronization. Moreover, by means of topological entropy, we show that these systems have better random-like behaviour, compared to conventional coupled map lattices (CML).

1. Introduction

The rich dynamics of chaotic systems is an attractive feature for many fields such as secure communications. Chaotic systems are, at the same time, deterministic and characterized by an unpredictable behaviour. Thus, from an information-theory viewpoint, they can be regarded as information sources [7]. Indeed, according to [10,11], a sequence of events conveys information if the events are not fully predictable.

The gap between the two fields, dynamical theory and communication theory, has been bridged thanks to symbolic dynamics [14]. Indeed, symbolic dynamics makes possible the conversion of chaotic trajectories into symbol sequences. It consists in partitioning the phase space into a finite number of disjoint partitions and in attributing a symbol to each one, so that, values falling in a given partition are presented by the same symbol.

The so obtained symbol sequences have shown to be good candidates in several communication applications such as in spread spectrum systems as pseudo-random codes [1-5].

For such applications, the chaotic sequences, used in the transmitter side of the communication system, have to be identically duplicated in the receiver side. Hence, chaotic systems, used in the transmitter and the receiver, have to be synchronized. The chaotic synchronization techniques which have been published to date are sensitive to both noise and distortion in the channel [1].

For this reason, we are interested in spatiotemporal chaotic systems for which synchronization is more efficient [4-6]. In particular, the spatiotemporal system that will be investigated in this paper shows better time synchronization than the conventional Coupled Map Lattices (CML).

Regarded as sources of information, we evaluate in this paper the amount of information generated by such systems, by means of topological entropy. Furthermore, we show that, besides the synchronization advantage, the proposed spatiotemporal system can generate sequences having better random-like behaviour, compared to the conventional CML.

In the second section of this paper, we propose a new family of spatiotemporal chaotic systems and show that it is more advantageous than the conventional CML, in terms of synchronization. Next, section III deals with topological entropy as a measure of unpredictability. First, we give the definition of this measure and its application in the low-dimensional chaotic systems. Then, we present its application in spatiotemporal systems and last we show and discuss simulation results. At last, some conclusions are drawn in section IV.

2. Synchronization of Spatiotemporal Chaotic Systems

The conventional Coupled Map Lattices (CML) are defined as follows [4-5]:

$$\begin{align*}
    x_i(k+1) &= (1-\varepsilon)f[x_i(k)] + \varepsilon f[x_{i-1}(k)] \\
    x_0(k) &= x(k)
\end{align*}$$

(1)

where $i$ is the space index, $i=1,\ldots, M$; $k$ is the time index, $k=1,\ldots,K$; $\varepsilon$ is the coupling coefficient. $f(.)$ is a one dimensional chaotic map, as for example the logistic map : $f(x) = 4x(1-x)$ or the piecewise-linear map : $f(x) = 4x \mod(1)$.

$x_{\alpha}(k)$ is the key sequence which is chosen to be a series of unformed distributed values in [0,1].

It has been shown in [4, 6] that the synchronization of two identical CML maps (1) having different initial conditions $x_{\alpha}(0)$ can be obtained, for some regions of values of the coupling strength $\varepsilon$, only by driving the two systems with the same key sequence. Moreover, we have shown in [6] that the behaviour of such a system, in terms of chaotic synchronization, depends on both $\varepsilon$ and $f$.

Now, let’s consider a new family of spatiotemporal chaotic systems defined by:
\[
\begin{cases}
 x_i(k+1) = (\alpha_i x_i(k) + \beta x_{i-1}(k)) \mod(1) \\
 x_0(k) = x(k)
\end{cases}
\]  

(2)

where \( \alpha_i \) and \( \beta \) are the system parameters. We choose in this study \( \alpha_0 = 0.005 \) and \( \beta = 4 \).

We point out that, owing to the \( \mod \) operator, the proposed CML do not diverge when all initial conditions are within \([0,1]\).

As the conventional CML (1), the synchrononization of this system can also be obtained by using the same key sequence.

Moreover, such a system can offer better chaotic-synchronizing time, compared to both logistic- and piecewise-based CML (1). In fig.1 we present the time needed for synchronizing each lattice site \( i \) \((N, M=100)\) for the new spatiotemporal system (2) and the conventional logistic- and piecewise-based CML (1) for a coupling strength \( \varepsilon = 0.97 \) (for which the CML shows the best behaviour [6]).

\[ L = \alpha \cdot \beta \]

\[ L \]

Consequently, the topological entropy takes its maximum value, which is [7-9]:

\[ h_T = \lim_{n \to \infty} \frac{\log P(n)}{n} = \log L \]

(3)

Chaotic systems can generate an infinite set of orbits by iterating different initial conditions, thanks to their unpredictable behaviour, which is characterized by sensitive dependence on initial conditions. These orbits can be converted into symbol sequences, by means of symbolic dynamics, which consists in partitioning the phase space and assigning a symbol to each partition [14]. Hence, chaotic systems can be considered as information sources generating symbol sequences [7]. The amount of information generated by such sources is then quantified by the topological entropy.

Thus, the greater the topological entropy, the greater the amount of unpredictability of the generated sequences and the better the random-like behaviour of the system.

For numerical computation, we consider a large set of \( n \)-symbols sequences generated as explained above, among which we compute the number of different sequences, say \( P(n) \). The topological entropy \( h_T \) is then [12-13]:

\[ h_T = \lim_{n \to \infty} \frac{\log P(n)}{n} \]

(4)

which can be approximated by a linear regression to the map \( \log P(n) \) versus \( n \), for \( n \) up to, say 16 [9].

We point out that the topological entropy depends on the chosen partition. However, we don’t address this topic in this paper.

1.2. Spatiotemporal Systems

Spatiotemporal systems have been investigated in secure communications for their advantageous behaviour, compared to low-dimensional ones [4-5].

For spatiotemporal systems, we generate symbol sequences in a slightly different way. For instance, for the system (1) we consider lattice sites \( \{x_i(1), x_i(2), \ldots, x_i(N)\} \) as orbits. Then, symbol sequences are obtained in the same way as low-dimensional chaotic systems, i.e. by means of symbolic dynamics. For instance we assign the symbol 0 or 1 to the values falling in \([0,0.5]\) or \([0.5,1]\), respectively.

The specificity, for such systems, is that each orbit depends not only on the initial condition \( (x_i(0)) \), but also on the previous orbit, through the coupling. Moreover, for some system parameters and coupling values, for which synchronization is achieved, each orbit depends only on previous one.

For \( i=1\ldots M \) and \( k=1\ldots N \), such a system generates a set of \( M \) sequences of length \( N \). We notice that \( M \) and \( N \) can
be as large as possible. Thus, the spatiotemporal system can be assimilated to a process generating $N$-symbol sequences. Consequently, the rate of information generated by such a system is given by the topological entropy, based on the computation of the number of different possible sequences. Thus, if the process is fully random, one could observe all $2^N$ possible sequences, in the case of two-symbol alphabet, for a given length $N$, and $M \geq 2^N$. In the ideal case, all these sequences are obtained for $M=2^N$.

1.3. Simulation Results

In the following we give a comparison between different systems: logistic- and piecewise-based CML (1), spatiotemporal system (2), according to the topological entropy.

The initial conditions and key sequences have been chosen the same for these comparisons.

In the first figure (fig. 2) we show the dependence of the approximated topological entropy $h_{Ta}$ on the coupling parameters.

We define $h_{Ta}$ as:

$$h_{Ta} = \frac{\log P(N)}{N}$$

(5)

$P(N)$ is the number of different $N$-symbol sequences among the set $M=2^N$ generated sequences.

- The piecewise CML shows entropy levels very close to the spatiotemporal system (2).
- We notice that the two first remarks confirm the results obtained in [6] concerning the correlation properties of the sequences generated by such systems. Indeed, we have shown that the best correlation levels (the lowest) are attained for high coupling values of the logistic-based CML, whereas, they maintain almost the same low level at all coupling values for the piecewise-based CML.

In figures 3 and 4, we present the asymptotic behaviour of $\log \frac{P(N)}{N}$ versus $N$, for $N$ up to 16, with $M=2^N$ and $M=2^{N+2}$, respectively. We choose $\varepsilon=0.97$ for the CML and $\beta=4$ for the spatiotemporal system (2). We can notice, for the two figures, that the proposed spatiotemporal system and the piecewise CML show the same asymptotic behaviour, hence the same topological entropy, whereas, the Logistic CML attain lower values than the two systems. This result is corroborated by the values of the topological entropy given in the table (1). These values have been obtained by linear regression of the map $\log \frac{P(N)}{N}$ versus $N$.

![Fig.2 $h_{Ta}$ versus coupling parameters ($N=10$).](image1)

![Fig.3 $\log \frac{P(N)}{N}$ versus $N$ for $M=2^N$.](image2)

![Fig.4 $\log \frac{P(N)}{N}$ versus $N$ for $M=2^{N+2}$.](image3)
When $M>2^N$ (Fig.4), the topological entropy is close to 1: the maximum value which can be attained with fully random process. Whereas, when $M=2^N$ (Fig.3), the topological entropy attains lower values. Consequently, the system conveys less information than the former case. This can be interpreted as follows: there is some duplicated sequences among the first $2^N$ generated ones. This duplication may be due to the choice of the partition or to the length of the sequences. Indeed, the duplicated sequences doesn’t originate in the same trajectories. Moreover, we can notice the increasing growth of $\log P(N)/N$ with $N$ in figure (3), which can be interpreted as follows: different trajectories may lead to symbol sequences having the same first sub-sequence up to some length.

In all cases, we can say that the spatiotemporal system (2) and the piecewise CML convey more information than the logistic CML. Hence, a better random-like behaviour of the generated sequences, in terms of unpredictability.

4. Conclusions

In this work, we have proposed a new family of spatiotemporal chaotic systems for the generation of symbol sequences. We have proved that these systems show better time synchronization than the conventional Coupled Map Lattices (CML). Furthermore, we have used topological entropy as a measure of unpredictability, to show that the proposed systems are comparable to the piecewise CML and better than the logistic CML, in terms of random-like behaviour.

Thus, the symbol sequences generated by such systems, show to be good candidates for secure communications and spread-spectrum systems, especially with the synchronization advantage.

References

Chaotic Synchronization in a Network of Living Mouse Pancreatic \(\beta\)-Cells

Takaya Miyano\(^\dagger\), Kensuke Ninomiya\(^\dagger\), Takahiro Kanno\(^\ddagger\), and Makoto Wakui\(^\ddagger\)

\(\dagger\)Department of Micro System Technology, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu, Shiga 525-8577 JAPAN

\(\ddagger\)Division of Blood Transfusion and Transplantation Immunology
School of Medicine, Fukushima Medical University
1 Hikariga-oka, Fukushima 960-1295, JAPAN

\(\ddagger\)Department of Physiology, Hirosaki University School of Medicine
5 Zaifucho, Hirosaki, Aomori 036-8562, JAPAN

Email: tmiyano@se.ritsumei.ac.jp

Abstract—This paper is concerned with tolerance in synchronous electrical activity of the pancreatic islet of Langerhans to inhomogeneities in the channel conductances of insulin secreting \(\beta\)-cells. Sherman’s \(\beta\)-cell model is utilized to simulate the electrical activity of living mouse pancreatic \(\beta\)-cells. Fidelity of numerical solutions to actual counterparts is assessed in terms of the permutation entropy of interspike time-intervals. Comparison of the permutation entropy between actual and numerical activity indicates that the statistical distribution in the channel conductances up to 15\% about their mean may be permissible to sustain actual synchronous activity.

1. Introduction

Pancreatic \(\beta\)-cell is a basic component of the islet of Langerhans as a micro-organ in the pancreas. Electrical characteristics of the \(\beta\)-cell have certain degrees of diversity, reflecting cell inhomogeneity. Single \(\beta\)-cells usually generate irregular spike trains of membrane potential. Surprisingly, a network of \(\beta\)-cells in the islet gives rise to synchronous bursting that consists of an alternating sequence of an active phase of spike generation followed by a quiescent phase [1], despite inhomogeneity of individual cells. Synchronous behavior of the islet is known to closely correlated with sensing glucose in the blood and subsequent insulin secretion [2, 3].

In this paper, we study tolerance in synchronous electrical activity of the pancreatic islet to inhomogeneities in the channel conductances of individual cells. To simulate actual electrical activity, we utilize Sherman’s model. There have been many propositions about \(\beta\)-cell modeling since the original work by Chay and Keizer [4]–[13]. Among others, Sherman’s model [5, 6] is useful to get the central idea of \(\beta\)-cell modeling. To assess the fidelity of numerical solutions to their actual counterparts, we resort to the permutation entropy of interspike time-intervals. The permutation entropy has been recently introduced by Bandt and Pompe to estimate a basic invariant measure of dynamical systems from time series [14]. \(\beta\)-cell activity can be viewed as a point process in which dynamical information is carried by a series of event timings. As for point processes, Sauer guaranteed that there is one-to-one correspondence between systems states and interspike interval (ISI) vectors under an appropriate embedding dimension [15, 16]. Based on estimates of the permutation entropy, we discuss the statistical distribution in the channel conductances that is permissible to sustain synchronous behavior of the islet.

2. Permutation Entropy

Permutation entropy is asymptotically equivalent to the Kolmogorov-Sinai entropy in the limit of infinite number of data points. See [14] for details. Let \(\pi = n!\) be all possible permutations of order \(n \geq 2\), representing coarse-grained patterns of dynamical behavior, for a sequence of \(n\) real numbers. Given a sequence of data points \(\{u_i\}_{i=0}^{N-1}\), we count the number of realizations of \(\pi\), denoting as \(m(\pi)\), for each embedding vector \(u = (u_1, u_{i+1}, \ldots, u_{i+n-1})\) of dimension \(n\). We calculate the relative frequency for \(\pi\) as \(p(\pi) = m(\pi)/(N-n+1)\) and define the permutation entropy as \(H(n) = -\sum_{\pi} p(\pi) \log_2 p(\pi)\). Bandt and Pompe pointed out that \(H(n)\) increases linearly with \(n\). Thus a practical redefinition of the permutation entropy is

\[
\tilde{h}(n) = \frac{H(n)}{(n-1) \log_2 n!}
\]

where \(0 \leq \tilde{h}(n) \leq 1\). The lower bound corresponds to a monotonous increasing or decreasing process, while the upper bound to a completely random sequence. Throughout this work, we utilize \(\tilde{h}(n)\) to represent the permutation entropy.
3. Physiological Data

A physiological experiment was conducted for NMRI mice under the approval for care and use of the animals by the ethical committee of Hirosaki University. For details of the experiment, see [1]. The membrane potential was observed using perforated patch-clamp technique for mice living β-cells perfused with 10 mM glucose. Figures 1 and 2 show electrophysiological records for a single cell and an intact cell of the islet during an active phase of bursting, respectively. Interspike time-intervals were estimated by reading times at which the voltage curve crossed upward the level 2/3 of the amplitude between the maximum and the minimum voltage. We thus obtained 30 and 39 ISI’s for the single cell and the islet, respectively.

Estimates of $\hat{h}(n)$ for the ISI data are shown in Fig. 3. The single cell yields smaller estimates than the intact cell of the islet. This implies more visible determinism in spike generation by the single cell, possibly deterministic chaos, which was confirmed in our previous work [17]. The aggregation of the cells in the islet causes an enhancement of complexity, despite the temporal order associated with synchronous activity.

Figure 1: Electrophysiological record for an isolated single β-cell.

Figure 2: Electrophysiological record for an intact β-cell of a pancreatic islet during an active phase.

4. β-Cell Model and Numerical Simulation

Sherman’s model is expressed as the following set of nonlinear ordinary differential equations.

$$\frac{dV}{dt} = -I_{Ca}(V) - I_K(V) - I_S(V)$$  (2)

where $I_{Ca}$ is the fast inward $Ca^{2+}$ current, $I_K$ the delayed-rectifier outward $K^+$ current, and $S$ a slow variable whose biological origin has been controversial. $V$ is the membrane potential, determined by the balance of the ionic currents. $n$ is the gating variable for the rectifier. The fast variables $V$ and $n$ generate electrical spikes, while $S$ is responsible for switching spike generation on and off during bursting. The conductances $g_{Ca}, g_K, g_S$ are dimensionless and scaled to remove the dependence on cell size, so that the equations include no capacitance.

We obtained numerical solutions for a single cell with the following set of parameters [5, 6]: $g_K = 10$, $g_S = 4$, $\tau = 0.02$ sec, $\tau_S = 35$ sec, $V_{Ca} = 25$ mV, $V_K = -75$ mV, $V_m = -20$ mV, $V_n = -16$ mV, $V_S = -45$ to $-25$ mV, $\theta_m = 12$ mV, $\theta_n = 5.6$ mV, $\theta_S = 10$ mV, $\lambda = 0.85$. The parameter $V_S$ works as a bifurcation parameter to the variable $S$ and is critical for structural instability of dynamical behavior of the β-cell. To find the optimal value of $V_S$, we extracted ISI’s from numerical sequences in much the same way as in the case of the physiological data, estimating...
\( \tilde{h}(n) \) as a function of \( V_S \). Results are shown in Fig. 4.Seemingly large variance at \( V_S > -39 \) mV reflects the bifurcation diagram of \( V_S \) versus \( S \). It is found that \( V_S = -38.34 \) mV provides the closest estimate to that of the actual single cell. This choice of \( V_S \) agrees with the previous works [7, 9].

Oscillatory bursting of the islet was simulated by chaotic synchronization of \( \beta \)-cells with mutual diffusive coupling[5]. The biological entity of the coupling has been considered to be gap junctions between neighboring \( \beta \)-cells. The coupling between the \( i \)th and the \( j \)th cell is introduced as

\[
\tau \frac{V_i}{dt} = -I_{\text{Ca}}(V_i) - I_K(V_i, n_i) - I_S(V_i, S_i) - g_c (V_i - V_j) \tag{11}
\]

where \( g_c \) is the dimensionless conductance. The simplest cell-cluster simulating the islet is a coupled pair of identical cells, where \( g_c = 0.1 \) according to Sherman’s prescription. The second simplest structure is an array of three identical cells. In this cluster, the central cell shares \( g_c = 0.05 \) with its neighboring cells. A more complex structure is a \( 5 \times 5 \) cell-lattice consisting of 125 identical cells. In this case, a central cell has its six nearest neighbors with which the total coupling conductance is shared. Thus, \( g_c \) is set to 0.1/6 \( \approx 0.017 \).

Numerical solutions of the homogeneous islet models were obtained at \( V_S = -38.34 \) mV. We obtained 32, 15, and 24 ISI’s for an active phase (consisting of about 3000 data points) of bursting of the 2-cell, 3-cell, and 125-cell model, respectively. To see how exactly each islet model mimics the actual islet, \( \tilde{h}(n) \) was estimated for the ISI data. Results are shown in Fig. 5. The 2-cell and 3-cell models failed in mimicking even the qualitative trend in the entropy relative to that of the single cell. In contrast, the 125-cell model successfully reproduced the actual trend. From these observations, the 125-cell model turns out to be a fairly good platform for simulating the islet.

![Figure 4: Normalized permutation entropy of numerical ISI's for a single cell as a function of \( V_S \) at the embedding dimension \( n = 3 \). A hundred thousands data points were used to generate ISI's.](image)

In order to simulate cell inhomogeneity, we next introduce statistical distributions to the channel conductances as:

\[
g_{\text{Ca}}(i) = 3.6 \times [1 + r \xi] \tag{12}
g_K(i) = 10 \times [1 + r \xi] \tag{13}
g_S(i) = 4 \times [1 + r \xi] \tag{14}
\]

where \( g_{\text{Ca}}(i), g_K(i), g_S(i) \) are the channel conductances of the \( i \)th cell, \( \xi \in N(0, 1) \) random real numbers of the Normal distribution, \( r \) a coefficient determining the variances: \( r = 0, 0.05, 0.1, 0.15, 0.2, 0.5 \). Figure 6 shows estimates of \( \tilde{h}(n) \) for actual and numerical ISI data (21, 21, 18, 20, and 23 ISI’s for \( r = 0.05, 0.1, 0.15, 0.2, \) and 0.5, respectively ). No synchronous bursting appeared when \( r > 0.5 \). The variances given by \( r = 0.1 \) and 0.15 are apparently able to reproduce the complexity of the actual dynamical behavior.

5. Discussion and Conclusion

The present findings have been acquired by applying a practical algorithm for estimating the Kolmogorov-Sinai entropy to the actual and numerical electrical activity as point processes. Inhomogeneities in the channel conductances relevant to ionic flows across the cell membrane are shown to be crucial to reproduce actual complex dynamical behavior. Such an effect would be difficult to directly observe by means of existing experimental methods, if not impossible. The tolerance in synchronous activity of the pancreatic islet to the differences in individual channel conductances is inferred to be 10–15% about their means.
In previous literature on modeling the pancreatic islet, authors often utilized cell-clusters of a size no larger than a few cells. However, the present work has demonstrated that the 2-cell model as well as the 3-cell model does not work at all in that it fails in reproducing even the qualitative tendency in the complexity of actual spike generation.

A remaining problem worth investigating might be heterogeneity in the structure of the islet. Actual islets consist of at least three distinct categories of cells, i.e., α-cells, β-cells, and δ-cells [1]. Each cell secretes distinct hormone, being organized into the islet in a complex fashion. Suppose that some α-cells and/or δ-cells are incorporated into a network of β-cells. These “alien” cells could be external perturbations to the β-cells to make some influence on spike generation. The pancreatic islet can be viewed as a complex network of the cells. It might be interesting to imagine a scale-free structure [18] or a highly optimized tolerance state [19] of the islet. Either structural characteristics could enhance the degree of complexity in spike generation.

In conclusion, the present work has shown that the permutation entropy can be a tractable and reliable tool to assess the complexity in spike trains of β-cells. From comparison of the estimates between actually observed activity and its numerical counterpart, we may predict the statistical distribution in actual channel conductances of individual cells to be 10–15% about their means.

Acknowledgments

This study was supported by grants-in-aid (No.15500126 to T.M.) from the Ministry of Education, Culture, Sports, Science and Technology.

References

Performance Analysis for Scheme of Synchronization-Manifold Shift Keying

Thang Manh Hoang and Masahiro Nakagawa

Department of Electrical Engineering, Faculty of Engineering, Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata 940-2188, Japan
Email: thang@pelican.nagaokaut.ac.jp, masanaka@vos.nagaokaut.ac.jp

1. Introduction

In recent years, synchronization of time-delay systems has been received much interest in applying in secure communication because a time-delay system generates a very highly dimensional dynamics and the number of positive Lyapunov exponents increases with the increasing in the length of time-delay [1]. More recently, lag synchronization of coupled multi-delay feedback systems has been aroused [2, 3]. The main difference of the synchronization of multi-delay feedback systems presented in Ref. 3 from those described in the other literatures is that the driving signal is in the form of multiple nonlinear transformations of delayed state variable. Accordingly, the chaotic secure communication system [4] has been designed by utilizing such synchronization model. So far, the communication system is said to be secure due to restriction of existing reconstruction methods in dealing with the synchronization of multi-delay feedback systems with multi-delay driving signal (MDFSMDDS) [7]. As presented in Ref. 4, the scheme of synchronization-manifold shift keying (SMSK) has been used for encoding various blocks of bits.

In this paper, we study the robustness of SMSK used in the proposed secure communication system by examining the average of synchronization errors as carrying and reference driving signals are transmitted through noise channels. The simulation result of specific example demonstrates the relation between the average of synchronization errors and the signal-to-noise ratio (SNR).

2. Chaotic Secure Communication System Using SMSK

In this section, lag synchronization of MDFSMDDS and the proposed secure communication system are reviewed.

2.1. Lag Synchronization of Coupled Multi-delay Feedback Systems

The dynamical equations for lag synchronization of coupled multi-delay feedback systems are in the form of:

\[
\text{Master}: \quad \frac{dx}{dt} = -\alpha x + \sum_{i=1}^{P} m_i f(x_{\tau_i}) \\
\text{Driving signal}: \quad DS(t) = \sum_{j=1}^{Q} k_j f(x_{\tau_{j+1}}) \\
\text{Slave}: \quad \frac{dy}{dt} = -\alpha y + \sum_{i=1}^{P} n_i f(y_{\tau_i}) + DS(t)
\]

where \(\alpha, m_i, n_i, k_j, \tau_i \in R\); integers \(P, Q (Q \leq P)\). \(x_t = x(t - \tau_i), y_t = y(t - \tau_i)\) are delayed state variables and denoted \(x_{\tau_i}\) and \(y_{\tau_i}\), respectively. \(f(.)\) is the differentiable generic nonlinear function. Note that, the driving signal is formed by sum of multiple nonlinearly transformed components of delayed state variable, and it is generated by a driving signal generator, denoted a DSG as in Fig. 1.

For the synchronization manifold \(y = x_{\tau_i}\), the derivative of synchronization error \(\Delta = y - x_{\tau_i}\) is

\[
\frac{d\Delta}{dt} = \frac{dy}{dt} - \frac{dx_{\tau_i}}{dt} = -\alpha \Delta + P \sum_{i=1}^{P} \left[ n_i f(x_{\tau_{i+1}} + \Delta_{\tau_i}) - (m_i - k_j) f(x_{\tau_{i+1}}) \right]
\]

where coefficient \(k_j\) belongs to transformation element of the driving signal, whose delay value satisfying \(\tau_{P+1} = \tau_d + \tau_i\). Also assumed the value relation of parameters be \(m_i - k_j = n_i, (1 \leq i \leq P \text{ and } 1 \leq j \leq Q)\), \(f(.)\) be bounded, and \(\Delta_{\tau_i}\) be small. Hence, Eq. (4) reduces to

\[
\frac{d\Delta}{dt} = \frac{dy}{dt} - \frac{dx_{\tau_i}}{dt} = -\alpha \Delta + P \sum_{i=1}^{P} n_i f'(x_{\tau_{i+1}}) \Delta_{\tau_i}
\]

where \(f'(.)\) denotes for the derivative of \(f(.)\). Applying the Krasovskii-Lyapunov theory [5, 6] to the case of multiple
delays, the sufficient condition for synchronization manifold is as below:

\[ \alpha > \sum_{i=1}^{P} |n_i \sup f'(x_{\tau_i + \tau_d})| \]

(6)

where \( \sup f'(\cdot) \) stands for the supreme limit of \( f'(\cdot) \).

### 2.2. Proposed Secure Communication System

For simplicity, the single-carrying-slave encoding model is considered and each block has \( l \) bits. As shown in Fig. 2, the encoder consists of one master, \( 2^l \) DSGs. The decoder has Carrying Slave, Reference Slave, \( 2^l \) forcing delay blocks \( d(\tau_{fs}), s = 1..2^l \), Error Estimator, and Synchronization Manifold Detector.

In operation, on one hand, the master synchronizes with carrying slave to generate multiple information-bearing synchronization manifolds (IBSM). On the other hand, the master synchronizes with reference slave to create reference synchronization manifold (RSM). \( IBSM_s \) and \( RSM \) are characterized by time delay \( \tau_{ds} \) and \( \tau_{dref} \), respectively. While each IBSM is used to encode a certain pattern of bit block, RSM is for detection the instantaneous IBSM. To transmit a certain block of bits, \( IBSM_s \) is created by synchronization between the master and the carrying slave in one synchronization period. For example, the blocks of 2 bits and assigned IBSMs are presented in Table 1. In the case of transmission of a number of blocks of bits, it is required a number of IBSMs to be generated consecutively, in other words, synchronization manifolds are switched among one another, i.e. modulation technique of SMSK.

Table 1: IBSMs and assigned blocks of bits

<table>
<thead>
<tr>
<th>( IBSM_1 )</th>
<th>( IBSM_2 )</th>
<th>( IBSM_3 )</th>
<th>( IBSM_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block of bits</td>
<td>00</td>
<td>01</td>
<td>10</td>
</tr>
</tbody>
</table>

As a result, the driving signal is merged by several parts of synchronization signals from DSGs. To identify \( IBSM_s \), the time delay \( \tau_{ds} \) must be detected by comparing averages of synchronization errors among one another at the end of every synchronization period to find the lowest one. Synchronization error \( \Delta_s(t) \) corresponding to \( IBSM_s \) is defined as \( \Delta_s(t) = y_{carr} - y_{ref} \), where \( y_{carr} \) and \( y_{ref} \) are state variables of carrying slave and that of reference slave after forcing delayed with the delay time \( \tau_{fs} \). The equation to estimate the average of synchronization errors is as follows:

\[ E_s = \frac{1}{T} \int_{t}^{t+T} \Delta_s^2(t)dt \]

(7)

where \( T \) is the time length of synchronization period. In a certain synchronization period, the lowest average of synchronization error is found by \( E_{lowest} = \min\{E_s; s = 1..2^l\} \) in Synchronization Manifold Detection. Moreover, to obtain the accurate detection, the relation between carrying manifold’s delay of IBSM \( \tau_{ds} \), reference manifold’s delay \( \tau_{dref} \), and forcing delay \( \tau_{fs} \) must be satisfied the equation of \( \tau_{ds} = \tau_{dref} + \tau_{fs} \). As a result, the ongoing \( IBSM_s \) is identified and the corresponding block of bits is recovered at the decoder side.

### 3. Robustness of Modulation Technique of SMSK

Note on the proposed secure communication system that there are two driving signals; one is for driving the carrying slave and the other is for synchronizing the reference slave. In this section, the operation of the proposed secure communication system is examined in the case that the driving signal is included by AWGN.
<table>
<thead>
<tr>
<th></th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
<th>$n_4$</th>
<th>$n_5$</th>
<th>$k_1$</th>
<th>$k_2$</th>
<th>$k_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carrying Slave</td>
<td>-0.5</td>
<td>-0.2</td>
<td>-0.8</td>
<td>-0.3</td>
<td>-0.7</td>
<td>-16.2</td>
<td>-17.5</td>
<td>-14.5</td>
</tr>
<tr>
<td>Reference Slave</td>
<td>-0.9</td>
<td>-0.2</td>
<td>-0.6</td>
<td>-0.3</td>
<td>-0.5</td>
<td>-15.8</td>
<td>-17.7</td>
<td>-14.7</td>
</tr>
</tbody>
</table>

Table 2: Value of parameters for simulation.

![Figure 4: Synchronization errors (a) without AWGN, (b) with SNR of 5(dB)](image)

Signals are transmitted through the additive white Gaussian noise channels as shown in Fig. 3. Specific example is illustrated with multiple delay Mackey-Glass systems and the simulation result shows the average of synchronization errors with respect to SNR.

With the presence of AWGN $\xi$, the driving signal in Eq. (2) becomes $DS(t) = \sum_{j=1}^{P} k_j f(x_{r_{t+j}}) + \xi$ and the derivative of synchronization error presented in Eq. (5) is rewritten as

$$\frac{d\Delta}{dt} = \xi - \alpha \Delta + \sum_{i=1}^{P} n_i f'(x_{r_{t+i}}) \Delta_{t_i}$$  \hspace{1cm} (8)

It is clear from Eq. (8) that the synchronization error vanishes to zero as the sufficient condition in Eq. (6) is satisfied, but it fluctuates around zero due to the presence of AWGN.

**Example 1:** To demonstrate the effect of AWGN on the synchronous regime, the synchronization of coupled five-delay feedback Mackey-Glass systems with and without AWGN is simulated. The dynamical equation is in the form of Eqs. (1)–(3) with $f(x) = \frac{\alpha x}{1 + x^3}$, and the value of parameters and delays is chosen as: $P = 5$, $Q = 3$, $b = 10$, $\alpha = 5.1$, $m_1 = -16.7$, $m_2 = -0.2$, $m_3 = -18.3$, $m_4 = -0.3$, $m_5 = -15.2$, $\tau_1 = 2.2$, $\tau_2 = 4.8$, $\tau_3 = 5.3$, $\tau_4 = 1.9$, $\tau_5 = 3.7$, $\tau_6 = 5.8$, $\tau_7 = 8.9$, $\tau_8 = 7.3$, $\tau_9 = 7.6$, $n_1 = -0.5$, $n_2 = -0.2$, $n_3 = -0.8$, $n_4 = -0.3$, $n_5 = -0.7$, $k_1 = -16.2$, $k_2 = -17.5$, $k_3 = -14.5$, $SNR = 5dB$. Fig. 4 shows the synchronization errors with and without AWGN. It is easy to observe that the synchronization error in the case of presence of AWGN as illustrated in Fig. 4(b) is significantly larger than that without AWGN as in Fig. 4(a).

**Example 2:** In this example, the robustness of SMSK is examined and the equations for the proposed secure communication system are in the form of five-delay Mackey-Glass as presented in Example 1. The value of parameters and delays is adopted as: $P = 5$, $Q = 3$, $b = 10$, $\alpha = 5.1$, $f(x) = \frac{\alpha x}{1 + x^3}$, $m_1 = -16.7$, $m_2 = -0.2$, $m_3 = -18.3$, $m_4 = -0.3$, $m_5 = -15.2$, $\tau_1 = 2.2$, $\tau_2 = 4.8$, $\tau_3 = 5.3$, $\tau_4 = 1.9$, $\tau_5 = 3.7$, $\tau_6 = 2.6$, $\tau_7 = 4.2$, $\tau_8 = 6.1$, $\tau_9 = 3.4$, $T = 20.0$. The blocks of 2 bits corresponding to IBSMs are as shown in Table 1. Tables 2 and 3 show the chosen value of parameters and delays, respectively. Table 4 presents the bit sequence “00011110011101001011” is broken into blocks of 2 bits and the appearance order of synchronization manifolds. As shown in Fig. 5, it is easy to identify the ongoing IBS $M_i$ in different synchronization periods by observing the synchronization error $\Delta_s(t)$.

As an exemplar case, a bit sequence is sent by the system with varying value of SNR, and means of squared synchronization errors (MSSE) is estimated for IBS $M_i$ corresponding a transmitted block of bits at a certain value of SNR. Note that AWGN in the carrying channel is independent from that in the reference channel, and the adopted
values of SNR are equal for both channels. The MSSE with respect to SNR of IBSM $s_i$ is illustrated in Fig. 6(a). It is clear to observe that MSSE decreases as SNR increases. As described above, an instantaneous IBSM is identified if the lowest average of synchronization error is found. With the presence of AWGN, the level of SNR affects on the accurate decoding should be analyzed numerically. The ratio of averages of synchronization errors is figured out as below:

$$R_{s,h} = \frac{E_{avr_s}}{E_{avr_h}}$$

where $E_{avr_s} = \frac{1}{N} \sum_{i=1}^{N} E_{s_i}$ and $E_{avr_h} = \frac{1}{N} \sum_{i=1}^{N} E_{h_i}$ are means of averages of synchronization errors corresponding to $s^{th}$ and $h^{th}$ synchronization manifolds, respectively. $N$ is the number of transmitted blocks of bits corresponding to IBSM $s_i$. For example, the ratios $R_{2,1}$, $R_{3,3}$ and $R_{2,4}$ are estimated in $N$ synchronization periods of IBSM $s_i$ corresponding to transmitted blocks of “01”. Inaccurate decoding results in case that the ratios for corresponding block of bits is nearly 1.0 at which the difference between averages of synchronization errors is so small to indicate instantaneous IBSM accurately. Figs 6(b)-6(e) show the ratio of averages of synchronization errors with respect to SNR, and the ratios decrease when SNR increases.

4. Conclusion

In this paper, the interference of AWGN on modulation technique of SMSK associated with the proposed secure communication system has been considered. The simulation result shows the robustness of SMSK. Moreover, the modulation technique of SMSK brings an interesting property by which transmission rate can be improved by expanding hardware, in other words, by increasing the number of IBSMs. In addition, SMSK technique can be used with other schemes of synchronization rather than the lag presented in this paper. As a result, it can be considered applying in practical chaotic communication systems.

Acknowledgments

This study was supported in part by the 21st Century COE (Center of Excellence) Program "Global Renaissance by Green Energy Revolution" and the Grant-in-Aid for Scientific Research (15300070) from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References


Complex Intermittency in Switching Converters

Yufei Zhou∗, Herbert H. C. Iu†, Chi K. Tse‡ and Jun-Ning Chen∗

∗School of Electronic Science and Technology, Anhui University, Hefei, Anhui, China
†School of Electrical, Electronic and Computer Engineering, The University of Western Australia, Australia
‡Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong

Abstract—Intermittent instability is commonly observed in switching power supplies during the design and development phase. It manifests as symmetrical period-doubling bifurcation in the time domain with long intermittent periods. Such intermittent operation is considered undesirable in practice and is usually avoided by appropriate adjustments of circuit parameters. This paper explores the mechanism and conditions for the emergence of intermittency in a common voltage-mode controlled buck converter.

1. Introduction

Intermittent operation, sometimes referred to as “breathing” in the physics literature [1]−[2], is a phenomenon which is frequently observed in periodically driven nonlinear systems. A specific form of intermittent operation exhibits period-doubling bifurcation in two symmetrical directions over the time domain. In this type of intermittent operation, the system intermittently bifurcates from the initial regular (or subharmonic) operation to the higher subharmonic operation or chaos, and then returns to the initial operation through the same bifurcation sequence in the reverse manner. To distinguish it from the usual bifurcation in the parameter space, we refer to this intermittent bifurcation in the time domain as time-bifurcation, as it manifests as a change of qualitative behavior of the system as time elapses [3] rather than as parameter changes. Such intermittency may arise in periodically driven nonlinear systems, where the frequency of a coupled signal is not consistent with the system’s driving frequency. Switching power converters are periodically driven nonlinear systems, to which abundant sources of periodic interference are coupled via unintended paths (e.g., conducted or radiated paths) [4].

When the interference frequency differs from the switching frequency, and if the interference is strong enough, intermittent operation occurs [5]. In practice, intermittency is an undesirable operating state that should be avoided because frequent deviations from the intended working regime may increase device stresses and hence jeopardize reliability.

This paper studies the simple voltage-mode controlled buck converter using a simple model that takes into account the coupling of intruding interference signal with the converter. With this model, the condition for the emergence of intermittency is studied in detail. Simulation results verify the analytical findings. The traditional bifurcation analysis deals with bifurcation in a parameter space, and hence is not directly relevant to the present study. By “transferring” the analytical procedure from a parameter domain to the time domain, we consider the variation of the characteristic multipliers as time elapses and establish a few important findings regarding the appearance of intermittent bifurcation in the voltage-controlled buck converter.

2. Voltage-Mode Controlled Buck Converter Coupled with Intruding Interference

2.1. Overview of Circuit Operation

The buck converter consists of an inductor, a switch, a diode, a storage capacitor and a resistor load, which are connected as shown in Fig. 1(a). When it operates under a common voltage-mode control, the buck converter has been shown to exhibit non-linear behavior [6]−[7]. Fig. 1(b) gives the key waveforms under this control. In practice, the intruding interference can take the form of coupling via conducted or radiated paths. Sometimes, the intruders can live on the same circuit board or be present at a very close proximity. Suppose the overall effect of the intruding signal is lumped to one spurious source $v_s$. We can model this coupling as an additive process which superposes the spurious signal $v_s$ directly on the control signal $V_{con}$ as shown in Fig. 1(a). The resulting perturbed control voltage is then given by

$$V_{con}^* = V_{con} + v_s$$  \hspace{1cm} (1)

When the circuit parameters are chosen as follows: $L = 20$ mH, $C = 47 \mu F$, $R = 22$ $\Omega$, $1/T = f_s = 2500$ Hz, $V_{ref} = 11.3$ V, $A = 8.4$, $V_{L} = 3.8$ V, $V_{C} = 8.2$ V, and without the interference ($v_s = 0$), the buck converter will experience a typical period-doubling bifurcation cascade with input voltage $V$ varying from 12 V to 33 V [3]. The first bifurcation occurs when $E \approx 24.6$ V, and the buck converter eventually enters a chaotic region when $E \approx 32.3$ V.

2.2. Response Induced by Weak Interference

With the same set of circuit parameters shown above, the unperturbed buck converter operates in a regular period-1 orbit (steady state) when $E$ is fixed at 22 V, which corresponds to $V_{con} \approx 6.2$ V. In general, if we consider the interference $v_s$ being a periodic signal, e.g., sinusoidal signal with amplitude $\hat{v}_s$, the perturbed control voltage can be written as

$$V_{con}^* = V_{con} + \hat{v}_s \sin(2\pi f_s t) = V_{con} [1 + \alpha_s \sin(2\pi f_s t)]$$  \hspace{1cm} (2)

where $f_s$ is the frequency of the interference, and $\alpha_s$ is the strength of the interference which is defined as the ratio of the $\hat{v}_s$ to $V_{con}$, i.e., $\alpha_s = \hat{v}_s / V_{con}$. Here, we assume that in the steady state, the ripple of $V_{con}$ is negligible. Coupling with this intruding interference signal, the converter will operate in a range of possible regimes, such as periodic operation, quasi-periodic operation, etc., according to the frequency ratio $\alpha_f$ ($\alpha_f = f_s / f_c$). We can summarize them in two cases.

Case 1: Frequency ratio $\alpha_f$ is an irrational number

In this case, the buck converter exhibits quasi-periodic behavior, because there are two incommensurate frequencies, i.e.,
Case 2: Frequency ratio and generates periodically with period number equalling \( N \). Poincaré section. Fig. 2 (a) shows an example with temperature. The quasi-periodicity is characterized by a torus on the switching frequency \( f_o \) waveform.

If \( \alpha_f = n + f/f_o \), i.e., \( \alpha_f = n + f/f_o \), where \( f \) is a small number compared to \( f_o \) and \( n \) is a rational number. This kind of interference will induce intermittency in a few different forms (such as intermittent subharmonics, intermittent chaos). The type and period of intermittency are determined by the interference frequency \( f_o \) and strength \( \alpha_f \).

When the interference frequency is close to the switching frequency, or its rational multiples, the perturbed control voltage \( V_{con} \) can be rewritten as

\[
V_{con} = V_{con}[1 + \alpha_f \sin 2\pi(nf_o + \tilde{f}t)].
\]

Assuming \( \tilde{f} = 1 \) and \( n = 1 \), we can obtain the time-bifurcation diagrams of intermittency, as shown in Fig. 3.

1. When the strength of the interference is very weak (i.e. small \( \alpha_f \)), the converter can still maintain its expected steady-state period-\( N_{den} \) operation, though the average operating point may fluctuate. The effect of the interference is not significant at this stage, and no intermittency shows up. Fig. 3(a) shows the corresponding time-bifurcation diagrams.

2. As the interference signal strength increases, the converter experiences higher subharmonic operation intermittently on top of the period-\( N_{den} \) operation. For a relatively low interference signal strength, period-2\( N_{den} \) subharmonics are observed intermittently with period-\( N_{den} \). Fig. 3(b) gives the corresponding time-bifurcation diagrams.

3. For a sufficiently high intruding signal strength, the converter starts to experience chaotic operation intermittently with period-2\( N_{den} \) subharmonics and the period-\( N_{den} \) operation, as shown in Fig. 3(c).

4. The intermittent period \( T_{int} \) is represented as \( T_{int} = \frac{1}{N_{den}/\alpha_f} \), where \( \tilde{f} = |f_o - nf_o| \). Thus, if the interference signal frequency is very close to the switching frequency of the converter, or its rational multiples, the intermittency is long. In the example here, \( T_{int} = 1 \) s with \( \tilde{f} = 1 \) Hz for the case of \( N_{den} = 1 \), which is shown in Fig. 3.

4. Transformation of Parameters

As the traditional bifurcation analysis mainly deals with parameter-bifurcations, it is not directly applicable to the analysis of intermittent bifurcation in the present study. We apply a transformation, which will transfer the change of time to the change of a new variable, “phase shift”, which is used to represent the equivalent drift in the interfering frequency from the switching frequency. Basically we express the perturbed control voltage, originally given in (4), as

\[
V_{con}' = V_{con}[1 + \alpha_f \sin 2\pi(nf_o + \tilde{f}t)]
\]

\[
= V_{con}[1 + \alpha_f \sin(2\pi nf_o t + \theta)]
\]

where \( \theta = 2\pi \tilde{f}t \). After this transformation, a sinusoidal term with the phase shift \( \theta \), and at the same frequency as the switching frequency \( f_o \), or its rational multiples, is obtained. Now we can map the change of the time over interval \([0, 1/\tilde{f}]\) onto the change of phase shift \( \theta \) over interval \([0, 2\pi]\).

3. Intermittency

As the intruding signal is coupled unintentionally, its frequency \( f_i \) is usually not equal to the switching frequency \( f_o \), or its rational multiples exactly. Therefore, it is unlikely to observe the kind of operation given in Fig. 2. But it is possible that the interference frequency \( f_i \) approaches the switching frequency \( f_o \), or its rational multiples. In this situation, we have \( f_i = nf_o + \tilde{f} \), i.e., \( \alpha_f = n + f/f_o \), where \( f \) is a small number compared to \( f_o \) and \( n \) is a rational number. This kind of interference will induce intermittency in a few different forms (such as intermittent subharmonics, intermittent chaos). The type and period of intermittency are determined by the interference frequency \( f_o \) and strength \( \alpha_f \).

Figure 1: Voltage-mode controlled buck converter coupled with spurious interference. (a) Schematic diagram; (b) key operation waveforms.
Using the above perturbed control voltage, a parameter-bifurcation diagram can be constructed by selecting $\theta$ as the bifurcation parameter, which is equivalent to the scenario of intermittency over $N_{\text{diff}}$ intermittent period. Using the same set of interference parameters as those used in Fig. 3, we obtain the corresponding parameter-bifurcation diagrams as shown in Fig. 4. When we compare the time-bifurcation diagrams and parameter-bifurcation diagrams, we find that they look very similar and the results are consistent.

5. Analysis of Intermittent Bifurcation

Essentially the parameter-bifurcation over the interval of $\theta$ is exactly the same as the time-bifurcation over an intermittent period shown in Figs. 3. In this section, we will first derive the iterative discrete-time map, and then analyze the stability of the converter by computing the Jacobian [8] and examining the loci of the characteristic multipliers [3,6].

In the case of $n = 1$, with the expected steady-state operation of the buck converter being a period-1 orbit, the converter will first lose stability and bifurcate from regular period-1 to period-2 subharmonic operation as $\theta$ increases. (Note that varying $\theta$ is equivalent to varying $t$, from (6), for a given period $1/f$.) A discrete-time iterative map (over one switching period)

$$x_{n+1} = f(x_n, \bar{d}_n)$$

should be constructed for the analysis of stability, where $d_n$ is the duty cycle and $\bar{d}_n = 1 - d_n$.

To complete the discrete-time model, we need to derive the defining function for the duty cycle. Essentially, we wish to find the connection between the switching instant $t_n$, or more precisely $\bar{d}_n$, and the state variables. We may define a switching function

![Figure 2: Poincaré sections of responses induced by different levels of interference with (a) $\alpha_f = \sqrt{3}/2, \alpha_i = 0.015$; (b) $\alpha_f = 1/2$ or $1/3$, $\alpha_i = 0.021$; (c) $\alpha_f = 1, \alpha_i = 0.0046$; (d) $\alpha_f = 2$ or $3$, $\alpha_i = 0.0046$.](image)

![Figure 3: Sampled inductor current waveforms (time-bifurcation diagram) for $f_s = n f_o + \bar{f} = f_o + 1 = 2501$ Hz. (a) $\alpha_i = 0.0046$; (b) $\alpha_i = 0.046$; (c) $\alpha_i = 0.069$.](image)

![Figure 4: Parameter-bifurcation diagrams with $\theta$ as bifurcation parameter for $n f_o = 2500$ Hz. (a) $\alpha_i = 0.0046$; (b) $\alpha_i = 0.046$; (c) $\alpha_i = 0.069$.](image)
With very weak interference ($\alpha = 0.0046$ for $nf_o = 2500$ Hz), the loci of the characteristic multipliers are found as in Fig. 5(a). From these figures, we observe that as the phase shift $\theta$ increases from 0, two characteristic multipliers move apart along a circle of radius less than 1 (equals 0.824 here) for the case of $nf_o = 2500$ Hz. When $\theta$ increases further, they move back along the original path. For all values of $\theta$, they stay within the unit circle, it implies that the converter maintains its expected operation. This result is consistent with the parameter-bifurcation diagrams shown in Fig. 4.

As the strength of the interference increases ($\alpha = 0.046$ for $nf_o = 2500$ Hz), bifurcation is observed as $\theta$ varies, as shown in Fig. 4(b), which corresponds to the intermittent subharmonics in Fig. 3(b). This scenario can be explained in terms of the movement of the loci of the characteristic multipliers, as illustrated in Fig. 5(b). For the case of $nf_o = 2500$ Hz, one characteristic multiplier crosses the unit circle along the real axis from outside when $\theta$ increases up to 1.6. Thus, the converter becomes stable as $\theta$ increases up to about 1.6. It remains in this stable operation state till $\theta = 5.25$. At this point, one characteristic multiplier leaves the unit circle and the converter undergoes a period-doubling.

For higher values of $\alpha$ (for $nf_o = 2500$ Hz), the converter period-doubles up to chaotic regime. The loci of the characteristic multipliers also locate the onset of the period-doubling bifurcation clearly. The movements of the characteristic multipliers are very similar to those in the previous case ($\alpha = 0.046$ for $nf_o = 2500$ Hz), and are omitted here.

6. Conclusion

In this paper we attempt to explore a commonly observed but rarely explained nonlinear phenomenon in power supplies. Sources of periodic interference are easily coupled to the converters via unintended paths. We find that when the interference frequency approaches the switching frequency or its rational multiples, intermittency occurs with the intended period-$N_{dcm}$ operation being interrupted. We also show that the signal strength and frequency of the interference signal determine the type and period of intermittency.

References


Experimental Confirmation of the Torus Birth through Border-collision Bifurcation in Voltage Controlled PWM-1 Buck Converter

Somnath Maity, Tapas K. Bhattacharya, and Soumitro Banerjee

Department of Electrical Engineering, I.I.T. Kharagpur-721302, India
E-mail: smaity, tapas, soumitro@ee.iitkgp.ernet.in

Abstract—The occurrence of quasiperiodicity in power electronic circuits have been observed in many circuit topologies, through smooth Neimark-Sacker bifurcation. Only recently, the direct transition from periodicity to quasiperiodicity through border collision bifurcation has been shown to occur in a multi-level converter. In this paper we show, both numerically as well as experimentally, that such an atypical bifurcation can occur in a very simple converter configuration: the voltage-mode controlled buck converter controlled by PWM-1 logic.

1. Introduction

Nonlinear phenomena of switched-mode power regulators have been studied extensively [1, 2]. It is well known that these circuits can be modeled as variable structure systems and give rise to a great variety of nonlinear behaviors—period doubling route to chaos, border collision bifurcation [3], grazing phenomena [4], quasi-periodicity, mode-locking behavior and torus breakdown [5]. Also, a route to quasi-periodicity through border collision bifurcation and its hysteresis phenomena has recently been studied in a multi-level converter, both numerically as well as experimentally [6].

In the present paper, we explore the occurrence of quasi-periodicity through border collision bifurcation in one of the simplest converter configurations: the voltage-mode controlled buck converter that uses PWM of type-1 (hereafter called PWM-1) [7, 8]. In the PWM-1 method, the control voltage at the beginning of the ramp cycle is compared with ramp waveform to generate the switching signal. This method of switching is suitable in digital control and VLSI implementation because it is necessary to sample the control signal only one per clock cycle. Moreover it automatically prevents the multiple switchings within the clock cycle.

This system consists of three subsystems, each one is chosen according to some switching conditions. There are five possible sequences of subsystems that can be followed within a clock cycle. This makes the discrete-time state space divided into five zones separated by four border lines. The derivation of exact functional form of those regions and the equations of separation lines will be reported elsewhere. There are five possible ways in which a normal period-one fixed point can hit a neighboring border line and can lose stability. Out of all the possible cases, we investigate one interesting behavior where there is a direct transition from period-one orbit to quasi-periodicity through border collision bifurcation when the converter moves from discontinuous conduction mode (DCM) to continuous conduction mode (CCM). Such mechanism of the birth of a torus was earlier reported in multilevel pulse-width modulated converters. In this paper we report experimental evidence of a direct transition from period-1 orbit to quasi-periodicity through border collision bifurcation in a PWM-1 controlled buck converter.

2. Discrete Model of the converter

Figure 1: PWM-1 voltage-mode controlled buck converter.

We consider a voltage-mode controlled buck converter as shown in Fig. 1. It consists of a controlled switch $S$ (MOSFET), an uncontrolled switch $D$ (diode), an inductor $L$, a capacitor $C$, and a load resistance $R$. The switching of the MOSFET is controlled by a feedback logic known as pulse width modulation of type-1 (PWM-1). This is achieved by obtaining a
control voltage \( v_{\text{con}} \), as a linear combination of the output capacitor voltage \( v_c \), and a reference signal \( V_{\text{ref}} \) in the form

\[
v_{\text{con}} = A(V_{\text{ref}} - v_c/k),
\]

where \( A \) is the gain of the error amplifier and \( k \) is the factor of reduction of the output voltage \( v_c \). An externally generated saw-tooth voltage \( v_{\text{ramp}} \), of time period \( T \) and upper and lower threshold voltages \( V_U \) and \( V_L \) respectively, is used to determine the switching instants. The value of the control voltage at the beginning of each ramp is compared with \( v_{\text{ramp}} \). The switch is turned on at the beginning of the clock period, and is turned off when \( v_{\text{con}} = v_{\text{ramp}} \). The inductor current ramps up during the on time, and falls during the off time. If the inductor current reaches zero value before the next clock cycle, it is discontinuous conduction (DCM) mode of operation.

There are three states of the system, described by three sets of differential equations, as follows.

**S0**: the equations for the on state

\[
\frac{d\bar{x}}{dt} = A_1 \bar{x} + B_1 V_{\text{in}}
\]

where

\[
A_1 = \begin{pmatrix} 0 & -1/L \\ 1/C & 1/RC \end{pmatrix}, \quad B_1 = \begin{pmatrix} 1/L \\ 0 \end{pmatrix}
\]

and \( \bar{x} = [i_L, v_c]^T \) and \( V_{\text{in}} \) is the input voltage.

**S1**: the equations for the off state

\[
\frac{d\bar{x}}{dt} = A_1 \bar{x}
\]

**S2**: the equations for the “discontinuous” state

\[
\frac{d\bar{x}}{dt} = A_0 \bar{x}
\]

\[
A_0 = \begin{pmatrix} 0 & 0 \\ 0 & -1/RC \end{pmatrix}
\]

The states are observed at every falling edge of the ramp signal to obtain a Poincaré map. Since the inductor current is zero in state S2, the system undergoes a change in dimension from 2-dimension to 1-dimension when the orbit changes from CCM to DCM.

Here, switching is modulated by the border function of each sub-system, i.e., switching occurs whenever the solution of each subsystem reaches the border specifically defined for that sub-system. The border function between the subsystem S0 and S1 is given by

\[
B_0 : \beta_0(x, t) = A \left( V_{\text{ref}} - \frac{v_c(t)}{k} \right) \left( \frac{V_U - V_L}{T} \right) t - V_L = 0;
\]

When \( \bar{x} \) hits the border \( B_0 \), switching occurs and subsequently the evolution of \( \bar{x} \) is governed by the subsystem S1. Two borders exist for the subsystem S1. One is the clock signal for resetting the switch that moves the system from S1 back to S0, and the other is at \( i = 0 \) which moves the system from S1 to S2.

These two borders can be described by

\[
B_{1a} : \beta_{1a}(x, t) = i_L = 0 \quad (6)
\]

\[
B_{1b} : \beta_{1b}(x, t) = t - nT = 0 \quad (7)
\]

While in S2, the system has only one border \( B_2 \) given by

\[
B_2 : \beta_2(x, t) = t - nT = 0
\]

If the state hits \( B_2 \), it returns to S0. Since in each ramp cycle the system may operate in CCM or in DCM, the model can be represented by the switching flow diagram shown in Fig. 2.

2.1. Bifurcation Analysis

2.2. Numerical Investigation

Fig. 3 shows that there can be five different ways in which the state can evolve from one clock instant to the next. As a result, the discretized state space can be divided into five compartments with five different mappings giving the complete model of the system. The stroboscopic map of the system is continuous throughout the state-space, but the derivatives are continuous only in the five regions and are discontinuous at the four borderlines. We explore the bifurcations that can occur when a fixed point moves across the border between DCM and CCM. The parameter set is chosen as: \( L=4mH, C=22\mu F, R=1380, A=11, k=6, V_{\text{ref}}=4.2 \text{ V}, V_U=5.5 \text{ V}, V_L=0.5 \text{ V}, \) and \( T=100\mu s\). The input voltage \( V_{\text{in}} \) is varied from 60 V to 30 V as the primary bifurcation parameter; to obtain the bifurcation diagram shown in Fig. 4.

It is found that a period-1 orbit abruptly changes to a period-4 orbit, where the system dimension undergoes a change from one to two as the orbit changes from DCM to CCM. For \( V_{\text{in}} \geq 56.47 \text{ V} \), the attractor is a period-1 DCM orbit, as succession of the states S0,
Figure 3: Possible evolutions of control voltage and inductor current between two clock instants where (a) control voltage $v_{\text{con}}|_{\text{at clock}} \geq V_r$; (b) $V_l < v_{\text{con}} < V_r$, and both on and off interval are included in the clock period; (c) $i$ becomes zero for a part of the clock cycle; (d) control voltage $v_{\text{con}}|_{\text{at clock}} \leq V_l$; (e) $i$ becomes zero for a part of clock cycle; (f) The critical case, where $i$ becomes zero just at the end of the clock period.

$S1$, and $S2$. By decreasing $V_{\text{in}}$, we observe the birth of a periodic orbit lying on a torus, and the transition occurs through a border collision bifurcation at $V_{\text{in}} \approx 56.46V$. The attractor is phase-locked, and the orbit contains some clock periods in CCM and some in DCM. Continuous-time plot of the orbit at the bifurcation point (not shown here for the sake of brevity) shows the inductor current reaching zero exactly at the end of a clock period, which confirms that the transition occurs through border collision bifurcation.

Using a numerical technique, we have calculated the eigenvalues of period-1 fixed point just before and after the border collision bifurcation. The eigenvalues of the Jacobian are 0 and -0.4271 for the parameter value $V_{\text{in}} = 56.47V$. For $V_{\text{in}} = 56.46$, there is a discontinuous jump of the eigenvalues. The fixed point becomes a spiral source as its characteristics multipliers are $\lambda_{1,2} = 0.2587 \pm 1.3651j$, their modulus being $|\lambda_{1,2}| = 1.3894 > 1$. The mathematical mechanism of such birth of torus and its subsequent evolution through resonance and ergodic tori have been investigated in [6].

Figure 4: Bifurcation diagram of PWM-1 voltage controlled buck converter taking $V_{\text{in}}$ as a bifurcation parameter.

2.3. Experimental confirmation

To verify these nonlinear phenomena, we have implemented the control logic experimentally. The experimentally obtained bifurcation diagram is shown in Fig. 5. The input voltage $V_{\text{in}}$ is taken as a bifurcation parameter. The diagram shows a transition from periodicity to quasi-periodicity at $V_{\text{in}} \approx 52.2V$, and as the parameter is further decreased, we observe a series of quasiperiodic and mode-locked behavior. The major mode-locked periodic windows also show a period-adding sequence.

Figure 5: Experimental bifurcation diagram of the PWM-1 voltage controlled buck converter.

Under nominal operating condition, converter operates in period-1 mode as shown in Fig. 6(a). As the parameter is decreased, the fixed point collides with the border between CCM and DCM at $V_{\text{in}} \approx 52.2V$, and the period-1 orbit abruptly changes to a quasiperiodic orbit (Fig. 6(b)). A new frequency component is born at this transition. Just after the bifurcation, the amplitude of the low-frequency component is small, which increases with further reduction of the parameter. Fig. 6(c-f) show the continuous-time waveforms
Figure 6: Experimental waveforms showing (a) normal period-1 operation at $V_{in} \approx 57.4 \text{ V}$, (b) just after the transition at $V_{in} \approx 52.2 \text{ V}$, (c) 1:7 mode-locked orbit at $V_{in} \approx 37.9 \text{ V}$, (d) quasi-periodicity with two incommensurate frequencies at $V_{in} \approx 37 \text{ V}$. (e) and (f) are the attractors in discrete phase space corresponding to (c) and (d).

and their corresponding discrete-time attractors. The discrete-time phase portraits are captured by taking the sampling the inductor current $i_n$ and control voltage $v_{con}$ in synchronism with the falling edge of the saw-tooth wave. The phase portrait shows a finite number of points lying on a closed loop for the mode-locked periodic orbits and a closed invariant curve when the waveform is quasiperiodic.

3. Conclusion

In this paper, we have demonstrated the transition from periodicity to orbits lying on a torus through border collision bifurcation in PWM-1 voltage controlled buck converter. We have calculated the eigenvalues of the Jacobian matrix before and after the border collision event and have shown that the eigenvalues discontinuously jump across the unit circle turning the fixed point abruptly into an unstable focus. Following this specific type of border collision bifurcation, a new low-frequency component is born in the system. The subsequent evolution of the system proceeds through interplay between the low frequency component and the clock frequency, as a sequence of mode-locked periodic windows and quasiperiodicity. To our knowledge, this is the simplest practical circuit in which this new type of bifurcation has been observed.

Acknowledgments

The work was partially supported by the Department of Atomic Energy, Government of India under project no. 2003/37/11/BRNS

References

Nonlinear Phenomena in Hybrid Replicator Dynamics with Changes of Interpretation Functions

Takafumi KANAZAWA† and Toshimitsu USHIO†

†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531 Japan,
Email: kanazawa@hopf.sys.es.osaka-u.ac.jp, ushio@sys.es.osaka-u.ac.jp

Abstract—We consider a situation that different populations interact with each other. Players of each population change their strategies according to payoffs and population states change dynamically. If some differences of perceptions arise between populations, then different strategies in a population may be perceived as indistinguishable ones in the other populations. Population’s perception is modeled by an interpretation function. Since populations’ perceptions are also affected by payoffs and/or population states, changes of interpretation functions must be taken into consideration. To discuss such a situation, hybrid replicator dynamics with changes of interpretation functions have been proposed. In this paper, we investigate nonlinear phenomena in hybrid replicator dynamics caused by switching between simple replicator dynamics. Especially, we consider that there exist two populations where one population has three pure strategies and the other has two. We introduce two parameters into each population’s payoff matrix, and we show how global properties of the model change depending on the parameters.

1. Introduction

The game theory has made considerable achievements in many scientific fields including economics, politics, and social science. Its extensions have been investigated in several different directions. Among them, we focus on an evolutionary game and a hypergame [1, 2].

In evolutionary game theory, the distribution of strategies in the population is changed according to payoffs which players earn depending on their selected strategies [3]. In this process, dynamical behaviors of the selection process are modeled by replicator dynamics [4, 5]. Replicator dynamics describes evolutions of the distribution of strategies in the population itself. On the other hand, in many conflict situations, players’ perceptions are different each other and their behaviors depend on their perceptions to the game. Such a situation is modeled by a hypergame [2, 6]. Various solution concepts and methods of analysis taking players’ perceptions into account have been proposed for the hypergames [7]. The hypergame theory investigates the effects of players’ perceptions about strategies and payoffs on their decisions.

To discuss dynamical behaviors of the selection process when taking into consideration players’ perceptions, replicator dynamics for multi-population hypergames has been proposed in [8]. A rigorous definition of player’s perception is given by an interpretation function which specifies the relationship between each player’s real strategies and strategies perceived by the other players. Replicator dynamics for hypergames is a model introducing the interpretation function into multi-population evolutionary games. In n-population models where each population has different perception, replicator dynamics with interpretation functions are formulated [8]. Moreover, as its extension, replicator dynamics with switching of the interpretation functions based on payoffs have been proposed [9]. Such a system is a hybrid system and its hybrid automaton representation is formulated. Moreover, using a well-known example of a hypergame “Soccer Hooliganism”, it has been shown that the orbits converging to heteroclinic orbits can appear by switching between two very simple dynamics.

In this paper, we investigate such nonlinear phenomena caused by switching between simple replicator dynamics. Especially, we consider that there exist two populations where one population has three pure strategies and the other has two. We introduce two parameters α and β into each population’s payoff matrix, and we show how global properties of the model change depending on the parameters.

2. Hybrid Replicator Dynamics

Suppose that there exist n large populations $P_1, P_2, \cdots, P_n$. For each population $P_i (i \in N = \{1, \cdots, n\})$, let $\Phi_i = \{1, 2, \cdots, m_i\}$ be a set of pure strategies of $P_i$, and $S_i$ be a set of population states of $P_i$, where a population state is a distribution of strategies in a population of players with a pure strategy. Let $\Phi_{ij}$ and $S_{ij}$ be $P_j$’s sets of pure strategies and population states perceived by $P_j$, respectively. Denoted by $s_j = (s_{i1}, \cdots, s_{im}) \in S_j$ is a population state of $P_j$, where $s_{ij}$ is the proportion of players with a pure strategy $k \in \Phi_j$ in $P_j$. We define sets of population state combinations $S$ by Cartesian product $S = \times_{i \in N} S_i$. For $s \in S$, denoted by $s_{-i}$ is the population state combination which results from $s$ by removing $s_i \in S_i$: i.e., $s_{-i} = (s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_n)$. Let $R_i : \times_{i \in N} S_j \to \mathbb{R}$ be the payoff function for population $P_i$, where $S_i = \mathbb{R}$. The mapping $f_{ij} : S_i \to S_{ij}$ is called $P_j$’s interpreta-
tion function to the population states of \( P_i \), where \( f_{ii} \) is the identity mapping: \( f_{ii}(s_i) = s_i \). We define \( P_j \)'s interpretation of a population state combination \( s \) by \( f_j(s) = (f_{j1}(s), \ldots, f_{jn}(s)) \). In each population, suppose that the rate of increase of players with a strategy \( k \) is expressed as the difference between the payoffs of a player with a strategy \( k \) and the average payoff of the population. Replicator dynamics with interpretation functions is given by

\[
\dot{s}_i^k = s_i^k \left[ R_i \left( f_i(s_{mi}, s_{-i}) \right) - R_i (f_i(s)) \right] = \xi_{i}^{f_i}(s),
\]

for all \( i \in N, k \in \Phi_i \), and the interpretation function \( f_i \), where \( e_i^k \) is the \( i \)-dimensional unit vector such that the \( k \)th element equals 1.

Let \( f(s) = (f_1(s), f_2(s), \ldots, f_n(s)) \in \mathcal{F} \) be an interpretation function combination, where \( f_i \in \mathcal{F}_i = \{f_{i1}^k, f_{i2}^k, \ldots, f_{in}^k\} \) is an element of a set of \( P_i \)'s interpretation of \( s \) and \( \mathcal{F} = \times_{i \in N} \mathcal{F}_i \) is the set of interpretation function combinations. Denoted by \( \sigma_{ik}^{f_2} \) is the event that \( P_i \) changes its interpretation function from \( f_1^k(s) \) to \( f_2^k(s) \). Let \( \Sigma \) be a set of all such events. We assume that the event \( \sigma_{ik}^{f_2} \in \Sigma \) occurs when \( P_i \) uses \( f_1^k(s) \) and the difference between the average payoff of \( P_i \) with \( f_1^k(s) \) and that with \( f_2^k(s) \) becomes equal or greater than a specified threshold \( \lambda_{ik}^{f_2} > 0 \).

According to the above definition, interpretation functions of each population change discretely, and evolutions of population states depend on replicator dynamics (1). Obviously, such a system is a hybrid system and we can describe our model as a hybrid automaton [10] by the following tuple [9]:

\[
RD = (\mathcal{F}, S, S_{ini}, S_{ini}, \Xi, \Sigma, G),
\]

where the set of interpretation function combinations \( \mathcal{F} \) corresponds to the set of discrete states, the set of population state combinations \( S \) corresponds to the set of continuous states, \( \mathcal{F}_{ini} \times S_{ini} \subseteq \mathcal{F} \times S \) is the set of initial states, \( \Xi = \{\xi_f : S \rightarrow \times_{i \in N} \mathbb{R}_{\geq 0}\}_{f \in \mathcal{F}} \) is the set of flows defining the continuous states, and \( G : S \rightarrow 2^{S} \) is the guard for each element of \( \Sigma \). For each \( f \in \mathcal{F} \), the flow \( \xi_f \in \Xi \) is defined by Eq. (1), where \( \xi_{f_i} = (\xi_{f_{i1}}^k, \ldots, \xi_{f_{in}}^m) \), and \( \xi_f = (\xi_{f_{11}}, \ldots, \xi_{f_{nj}}) \). Moreover, for an event \( \sigma_{ik}^{f_2} \in \Sigma \), the guard \( G \) is defined as follows:

\[
G(\sigma_{ik}^{f_2}) = \left\{ s \in S \mid R_i \left( f_i^k(s) \right) - R_i \left( f_i^k(s) \right) \geq \lambda_{ik}^{f_2} \right\}.
\]

and an event is assumed to occur whenever the guard for the event is satisfied.

When an event occurs, if the guard for an another event is satisfied at that instant in the new discrete state, the event occurs without changes of the continuous state. However, an infinite chain of events without changes of the continuous state cannot exist since we assume \( \lambda_{ik}^{f_2} > 0 \) in Eq. (3). By changing of the interpretation functions, the average payoff of the population must increase. So it is impossible to visit the same discrete state more than once by a sequence of events without changes of the continuous states.

### 3. Nonlinear Phenomena in Hybrid Replicator Dynamics

The model of hybrid replicator dynamics has been applied to a well-known example called “Soccer Hooliganism” and shown that behaviors converging to heteroclinic orbits can appear in it [9]. However, its mechanism has not been discussed. So, we investigate nonlinear phenomena caused by switching between simple replicator dynamics shown in [9].

Suppose that there exist two large populations \( P_1 \) and \( P_2 \). \( P_1 \) has three pure strategies and \( P_2 \) has two. We consider two types of interpretation function combinations \( f^1 = (f_1, f_2) \) and \( f^2 = (f_1, f_2) \), where \( f_1(s) = (s_1, s_2) \), \( f_2(s) = (s_1^2 + s_1^3, s_1) \) and \( f_3(s) = (s_1^2 + s_1^3, s_1) \). Let us suppose that it is possible for \( P_2 \) to change the interpretation functions from \( f^1(s) \) (resp. \( f^2(s) \)) to \( f^3(s) \) (resp. \( f^2(s) \)). Hence, \( \Sigma = \{\sigma_{12}^{f_{2}}, \sigma_{21}^{f_{2}}\} \). Additionally, let \( A = (a_{ij}) \), and \( B = (b_{ij}) \) be \( P_1 \)'s and \( P_2 \)'s payoff matrices, respectively, where \( R(e_i, e_j) = a_{ij} \) and \( R(e_i', e_j') = b_{ij} \). We defined them as follows:

\[
A = \begin{bmatrix} 3 & 4 \\ 2 & 5 \\ 1 + \alpha & 6 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 2 & 2 + \beta \\ 4 & 2 \end{bmatrix},
\]

where \(-1 < \alpha < 1 \) and \(-1 < \beta < 1 \) are parameters.

Note that \( \sum s_1 = \sum s_2 = 1 \). So, we describe a population states combination by \( s = [s_1, s_1', s_1''] \). From Eq. (1), replicator dynamics is formulated as follows:

\[
\dot{s}_1 = s_1 \left[ (4 - \alpha) s_1 - 2 \right] - \frac{s_1^2}{2} \left[ (2 - \alpha) s_1^2 - 1 \right],
\]

\[
\dot{s}_2 = s_2 \left[ (4 - \alpha) s_2 - 2 (1 - \alpha) s_2^2 \right] - \frac{s_2^2}{2} \left[ (2 - \alpha) s_2^2 - 1 \right],
\]

\[
\dot{s}_1' = \left\{ \begin{array}{ll}
2 s_1' (1 - s_1') (\beta - 2 s_1') & \text{for } f^2_1,
\end{array} \right.
\]

\[
\dot{s}_2' = \left\{ \begin{array}{ll}
2 s_2' (1 - s_2') (\beta - 2 s_2') & \text{for } f^2_2.
\end{array} \right.
\]

From condition (3), the guard is given as follows:

\[
G(\sigma_{12}^{f_{2}}) = \left\{ s \in S \mid -g(s) \geq \lambda_{12}^{f_2} \right\},
\]

\[
G(\sigma_{21}^{f_{2}}) = \left\{ s \in S \mid g(s) \geq \lambda_{21}^{f_2} \right\},
\]

where

\[
g(s) = R_2 \left( f_2^1(s) \right) - R_2 \left( f_2^2(s) \right) = s_1^2 \left[ 2 - (\beta + 2) s_1^2 \right].
\]

If \( \alpha = 0 \) and \( \beta > 0 \) holds, then behaviors converging to heteroclinic orbits appear similar to the simulation in [9] as shown in Fig. 1. By using parameters \( \alpha \) and \( \beta \), we investigate mechanisms of such nonlinear phenomena.

Table 1 shows vector fields on the boundary edges, equilibrium points on each interpretation function, and contours of \( g(s) \).

(i) The case \( \beta > 0 \): On each discrete state \( f^1 \) and \( f^2 \), there exist equilibrium points different from the vertices.
Table 1: Vector fields on the boundary edges, equilibrium points on each interpretation function, and contours of $g(x)$.

<table>
<thead>
<tr>
<th>$\alpha &lt; 0$</th>
<th>$\alpha = 0$</th>
<th>$\alpha &gt; 0$</th>
<th>Contour of $g(s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta &lt; 0$</td>
<td>$\beta = 0$</td>
<td>$\beta &gt; 0$</td>
<td></td>
</tr>
</tbody>
</table>

- **$\alpha < 0$:** There exist isolated equilibrium points $[0, 0, \frac{\beta}{\pi^2}, \frac{2}{\pi^2} s_1, \frac{2}{\pi^2} s_2, \frac{1}{\pi^2} s_1]$ on both $f^1$ and $f^2$, and $[0, \frac{\beta}{\pi^2}, \frac{2}{\pi^2} s_1, \frac{1}{\pi^2} s_1, \frac{1}{\pi^2} s_2]$ on $f^1$, and $[0, \frac{\beta}{\pi^2}, \frac{2}{\pi^2} s_1, \frac{1}{\pi^2} s_2]$ on $f^2$. All orbits on $s_1^2 = 0$, $s_1^1 = 0$ on $f^1$, and $s_1^1 + s_1^2 = 1$ on $f^2$ are nonisolated periodic orbits.

- **$\alpha = 0$:** There exist nonisolated equilibrium points $[s_1^1, \frac{\beta}{\pi^2}, s_1^1 - s_1^2, \frac{1}{\pi^2}]_1$ on $f^1$ and $[\frac{\beta}{\pi^2}, s_1^1, \frac{1}{\pi^2}]_1$ on $f^2$. On each discrete states $f^1$ and $f^2$, though the orientations of the vector fields on the boundary edges in the cases $\alpha < 0$, $\alpha = 0$, and $\alpha > 0$, are the same as shown in Table 1. Since the magnitude of the vectors are different from each other, however, global properties of these cases are different from each other.

On each discrete states $f^1$ and $f^2$, though the orientations of the vector fields on the boundary edges in the cases $\alpha < 0$, $\alpha = 0$, and $\alpha > 0$, are the same as shown in Table 1. Since the magnitude of the vectors are different from each other, however, global properties of these cases are different from each other.

- **$\alpha > 0$:** Orbits on both $f^1$ and $f^2$ converge to the plain $s_1^2 = 0$.

In hybrid replicator dynamics, by switching between replicator dynamics whose properties are explained above, the following orbits exist depending on the value of $\alpha$:

- **$\alpha < 0$:** Orbits which converge to the heteroclinic orbit $[0, 0, 0] \rightarrow [0, 0, 1] \rightarrow [0, 1, 1] \rightarrow [1, 0, 1] \rightarrow [1, 0, 0] \rightarrow [0, 0, 0]$. Figure 2 shows an example of such an orbit.
Figure 3: Example of orbits in the case $\alpha > 0$ and $\beta > 0$

Figure 4: Example of orbits in the case $\alpha > 0$ and $\beta = 0$

- $\alpha > 0$: Orbits which converge to the plane $s_{1}^{2} = 0$. Figure 3 shows an example of such an orbit.

- $\alpha = 0$: Orbits which converge to the heteroclinic orbit consists of the orbits on the edge $[0, 0, 0] \rightarrow [0, 0, 1]$, the orbits from $[0, 0, 1] \rightarrow [1, 0, 1]$ on the plain $s_{1}^{2} = 1$, on the edge $[1, 0, 1] \rightarrow [1, 0, 0]$, and the orbits from $[1, 0, 0]$ to $[0, 0, 0]$ on the plain $s_{1}^{2} = 0$. Unlike the case $\alpha \neq 0$, however, the orbits on $s_{1}^{2} = 0$ (resp. $s_{1}^{2} = 1$) do not converge to the boundary edges $[1, 0, 0] \rightarrow [0, 0, 0]$ nor $[1, 0, 1] \rightarrow [0, 1, 0] \rightarrow [0, 0, 0]$ (resp. the edge $[0, 0, 1] \rightarrow [1, 0, 1]$ nor $[0, 0, 1] \rightarrow [0, 1, 1] \rightarrow [1, 0, 1]$).

With decrease of $\beta > 0$, the equilibrium points which are not on the vertices of $\mathcal{S}$ approach to $s_{1}^{2}$ axis or the edge $s_{1}^{1} = 0$, $s_{1}^{2} = 1$. In the case $\beta = 0$, they are to be on the axis or the edge as explained next.

(ii) The case $\beta > 0$: Any points on $s_{1}^{2}$ axis and $s_{1}^{1} = 0$, $s_{1}^{2} = 1$ on $f_{2}^{\sigma}$ became nonisolated equilibrium points. Moreover, since the equilibrium points which are not on the vertices in the case $\beta > 0$ are to be on $s_{1}^{2}$ axis or $s_{1}^{1} = 0$, $s_{1}^{2} = 1$ on $f_{2}^{\sigma}$, the periodic orbits which exist in the case $\beta > 0$ disappear. Each orbits starting from the interior of $\mathcal{S}$ converge to one of the points on $s_{1}^{2}$ axis depending on its initial value. Additionally, since $g(s)$ is always positive in the case $\beta \leq 0$ as shown in Table 1, the condition $s \in G(\sigma_{1}^{12})$ can not be satisfied for all $x_{1}^{12} > 0$. So the event $\sigma_{1}^{12}$ can not occur, and $P_{2}$’s interpretation function can not be switched from $f_{1}^{12}$ to $f_{2}^{12}$. Figure 4 shows an example of orbits in the case $\beta = 0$.

(iii) The case $\beta < 0$: The orientation of the vector field on $s_{1}^{2}$ axis and the line $s_{1}^{1} = 0$, $s_{1}^{2} = 1$ on $f_{2}^{\sigma}$ are opposite of that for $\beta > 0$, i.e., they are negative. So, the point $[0, 0, 0]$ is the unique asymptotically stable equilibrium point and all orbits starting from the interior of $\mathcal{S}$ converge to that point regardless of switching of interpretation functions.

4. Conclusions

In this paper, we have investigated nonlinear phenomena in hybrid replicator dynamics with changes of interpretation functions. Especially, we consider that there exist two populations where one population has three pure strategies and the other has two. We have introduced two parameters $\alpha$ and $\beta$ into each population’s payoff matrix, and we have shown how global properties of the model change depending on the parameters. It is future work to discuss stability conditions for interpretation functions and population states to converge to an equilibrium point by using stability theory of hybrid systems.

References


Fractals in Hybrid Systems

Kazutoshi Gohara

Graduate School of Engineering, Department of Applied Physics, Hokkaido University, Sapporo 060-8628, Japan
Email: gohara@eng.hokudai.ac.jp

Abstract—In this paper the hybrid dynamical system is defined by a continuous dynamical system discretely switched by external temporal inputs. A theory developed by the author suggests that the dynamics of ordinary differential equations, which is stochastically excited by external temporal inputs, is characterized by a set of continuous trajectories with a fractal structure in hyper-cylindrical phase space. After the theory is reviewed, a simple example and a preliminary multi-fractal formalism are given.

1. Introduction

A Hybrid Dynamical System (HDS) [1] is a dynamical system that involves an interaction of discrete and continuous dynamics. Control strategies introduced for HDS have been applied to various design problems for realizing desired behavior in power plants [2], chemical plants [3], etc.

The dynamical evolution of HDS can be described by a differential equation involving a number of vector fields that are switched one after another [4]. It has been shown that HDS displays very complex behaviors such as chaotic behavior[5]. Branicky [6] and the author [7] have independently demonstrated numerical experiments of HDS described by simple linear equations and shown that the state of the system moves around on the Sierpinski gasket, a very well-known fractal set. These results suggest that the fractals may universally appear in some classes of HDS. In this paper a theory for continuous dynamical systems with temporal inputs are presented from hybrid dynamical systems point of view.

We focus on dissipative, continuous, and non-autonomous dynamical systems defined by the following ordinary differential equations:

$$\dot{x} = f(x, t), \quad x \in \mathbb{R}^N,$$  

where $x$, $t$ and $f$ are state, time, and vector field, respectively. Equation (1) implies that the vector field depends on time. In general, this suggests that a system is influenced by other systems. To emphasize that the vector fields depend on time throughout the input $I(t)$, we rewrite Eq. (1) as follows:

$$\dot{x} = f(x, I(t)), \quad x, I \in \mathbb{R}^N.$$

2. Dynamics with Periodic Inputs

We will begin by considering a dynamics with a periodic input:

$$I(t) = I(t + T),$$

where $T$ is the period of the input. The vector field $f$ is also periodic with the same period $T$:

$$f(t) = f(t + T).$$

Introducing the angular variable $\theta = \frac{2\pi}{T} t$ mod $2\pi$ and new state variable $y = (x, \theta)$, we can transform the non-autonomous system expressed by Eq. (2) into the following autonomous system:

$$\dot{y} = f(y), \quad y \in \mathbb{R}^N \times S^1.$$  

The vector field $f^*_1$ is defined on a manifold $M : \mathbb{R}^N \times S^1$ that is a hyper-cylindrical space. In other words, Eq. (3) expresses a continuous dynamical system $D_c$ defined by the manifold $M$ and the vector field $f_1$:

$$D_c = (M, f_1).$$

In the hyper-cylindrical space $M$, we can define the Poincaré section:

$$\Sigma = \{(x, \theta) \in \mathbb{R}^N \times S^1| \theta = 2\pi\},$$

where a trajectory starting from an initial state at $\theta = 0$ returns at $\theta = 2\pi$. On the section $\Sigma$, a mapping can be defined which transforms a state $x_\tau$ to another state $x_{\tau+1}$ after interval $T$:

$$x_{\tau+1} = g_I(x_\tau), \quad x_\tau \in \mathbb{R}^N,$$  

where $g_I$ is an iterated function. In other words, Eq. (5) expresses a discrete dynamical system $D_d$ defined by the manifold $\Sigma$ and the iterated function $g_I$:

$$D_d = (\Sigma, g_I).$$

We can summarize the dynamics with a periodic input as follows. The periodic input $I$ defines two dynamical systems, a continuous one $D_c$, and a discrete one $D_d$, defined by Eqs. (4) and (6), respectively. In order to emphasize the relation among $I, D_c$ and $D_d$, we use the following schematic expression:

$$I \rightarrow D_c \rightarrow D_d.$$
3. Dynamics with Switching Inputs

3.1. A Set of Inputs

In this section, we consider a dynamics in which plural input patterns are stochastically fed into the system one after the other. Let us suppose that each input is one period of a periodic function. For example, we can define the periodic function by the following Fourier series:

\[ I(t) = \frac{a_0}{2} + \sum_{m=1}^{M} \left( a_m \cos \frac{2\pi m}{T} t + b_m \sin \frac{2\pi m}{T} t \right) , \]

where \( a_0, a_m, b_m \in \mathbb{R}^N \) are vectors for Fourier coefficients, and \( T \) is the period. The set of these parameters defines the input space:

\[ \mathcal{I} = \left\{ a_0, (a_m, b_m)_{m=1}^{M}, T \right\}, \]

\[ \mathcal{I} : \mathbb{R}^{N+2N(M+1)}. \]

Within this space, an arbitrary point represents an external temporal input. We consider the input as a set \( \{I_i\} \) of time functions \( I_i \) sampled on the parameterized space \( \mathcal{I} \). In the following sections, we abbreviate subscripts and express individual sets as \( [\cdot] \) for simplicity.

3.2. Two Sets of Dynamical Systems

Much as in the case of periodic input, we can define two sets of dynamical systems corresponding to the set \( \{I_i\} \). One is the set of continuous dynamical systems:

\[ \{D_{ci}\} = (\mathcal{M}, \{f_i\}), \]

and \( \{f_i\} \) is the set of vector fields on the hyper-cylindrical space \( \mathcal{M} \). The other is the set of discrete ones:

\[ \{D_{di}\} = (\mathcal{\Sigma}, \{g_i\}), \]

where \( \{g_i\} \) is the set of iterated functions on the global Poincaré section \( \mathcal{\Sigma} \). We also use the following schematic expression, which is similar to expression (7):

\[ \{I_i\} \rightarrow \{D_{ci}\} \rightarrow \{D_{di}\}. \]

3.3. Excited Attractor

In this paper, we are considering a continuous dynamical system that is dissipative and has an attractor for a periodic input. When an input pattern is fed into the system repeatedly, i.e., in the case of periodic input, a trajectory converges to an attractor. But how do we describe the dynamics when the inputs are switched stochastically? Even for an input with finite interval, we can assume an attractor corresponding to a periodic input with infinite interval. We call this an excited attractor in order to emphasize that the attractor is excited by the external input. Although a trajectory tends to converge to a corresponding excited attractor, the trajectory cannot reach the excited attractor due to the finite time interval. If the next input is the same as the previous one, the trajectory again goes toward the same excited attractor. If the next input is different from the previous one, the trajectory changes its direction and goes toward an excited attractor distinct from the previous one. Continuing this process, the trajectory takes a transient path to the excited attractor. Intuitively, the trajectory will be spread out around excited attractors in the hyper-cylindrical phase space \( \mathcal{M} \). How, then, do we characterize the properties of the transient trajectory?

4. Fractal Transition

4.1. Iterated Function System

In the following two sections, we focus on the set \( \{g_i\} \) of iterated functions on the global Poincaré section \( \mathcal{\Sigma} \).

4.1.1. Hutchinson’s Theory

Hutchinson [8] has proved that a set \( \{h_i\} \) of iterated functions, which are not limited on the Poincaré section, defines a unique and invariant set \( C \) that satisfies the following equation:

\[ C = \bigcup_{i=1}^{L} h_i(C), \]

where

\[ \bigcup_{i=1}^{L} h_i(C) = h_1(C) \cup h_2(C) \cup \cdots \cup h_L(C), \]

and

\[ h_i(C) = \bigcup_{x \in C} h_i(x). \]

Such an invariant set \( C \) is often a fractal or sometimes used as a mathematical definition of various fractals.

A sufficient condition for satisfying Eq. (12) is the contraction property of \( h_i \) for all \( i = 1, 2, \ldots, L \). The contraction for \( h_i \) is definitely defined by the Lipschitz constant \( Lip(h_i) \):

\[ Lip(h_i) = \sup_{x \neq y} \frac{d(h_i(x), h_i(y))}{d(x, y)}, \]

where \( d \) is a distance on a metric space. When

\[ Lip(h_i) < 1, \]

the map \( h_i : x \rightarrow y \) is called the contraction or the contraction map.

4.1.2. Iterated Function System with Probabilities

Barnsley has named a set \( \{h_i\} \) as the IFS (Iterated Function System) [9]. He introduced the IFS with probabilities as follows:

\[ (h_i, \{p_i\}), \]

where \( \{p_i\} \) is a set of probabilities corresponding to \( \{h_i\} \).
Based on the IFS with probabilities, he proposed an alternative method for constructing the invariant set $C$ that satisfies Eq. (12). The iterated functions $h_l$ are switched with probabilities $p_l$ for $l = 1, 2, \ldots, L$ as follows. Choose an initial point and then choose recursively and independently $x_{\tau+1} \in \{ h_1(x_\tau), h_2(x_\tau), \ldots, h_L(x_\tau) \}$ for $\tau = 0, 1, 2, \ldots, \infty$, where the probability of the event $x_{\tau+1} = h_l(x_\tau)$ is $p_l$. Thus a sequence constructs a set $\{x_{\tau}\}_{\tau=0}^{\infty}$. Using Hutchinson’s theory, Barnsley has shown that the set $\{x_{\tau}\}_{\tau=0}^{\infty}$ constructed by random sequence, and here assumed to have equal probability, “converges to” the set $C$ defined by Eq. (12) when all iterated functions are the contractions. The set $\{x_{\tau}\}_{\tau=0}^{\infty}$ is thus an approximation of $C$.

4.2. Vector Field System

We are now ready to consider the trajectory of continuous dissipative dynamical systems excited by the temporal inputs. When the inputs $I_l$ are stochastically fed into the system one after another, the vector fields $f_l$ and the iterated functions $g_l$ are also stochastically switched as explained in Sec. 3. To emphasize the relation among the set $\{I_l\}$, $\{f_l\}$ and $\{g_l\}$, we use the following schematic expression instead of expression (11):

$$I_l \rightarrow f_l \rightarrow g_l.$$  \hspace{1cm} (15)

We call the set $\{f_l\}$ the Vector Field System (VFS), which is similar to the Iterated Function System (IFS) for the set $\{g_l\}$. The discrete dynamics on the Poincaré section $\Sigma$ correspond to the random iteration algorithm using the IFS with probabilities. That is, when all iterated functions $g_l$ are the contractions, the state $x_\tau$ on the Poincaré section approximately changes on the invariant set $C$ defined by Eq. (12) after sufficient random iterations. The property of the set $C$ having the fractal structure affects the trajectory in the hyper-cylindrical phase space $\mathcal{M}$.

The trajectory set $\Gamma(C)$ corresponding to the input set $\{I_l\}$ is obtained by the union of the trajectory set $\gamma_l(C)$ for each input $I_l$:

$$\Gamma(C) = \bigcup_{l=1}^{L} \gamma_l(C),$$

$$= \gamma_1(C) \cup \gamma_2(C) \cup \cdots \cup \gamma_L(C).$$  \hspace{1cm} (16)

We can conclude that the dissipative dynamical systems excited by plural temporal inputs are characterized by the trajectory set $\Gamma(C)$ starting from the initial set $C$ defined by Eq. (12). All of the trajectories are considered to represent the transition between the excited attractors. We call this the fractal transition between the excited attractors. At this point, we should emphasize that the contraction property of iterated functions defined on the Poincaré section is a sufficient, but not necessary, condition for the fractal transition.

5. Simple Example

In this section a simple example is shown using the solvable class of the following two dimensional linear ODE with additive input [7].

$$\dot{x} = Ax + I(t),$$  \hspace{1cm} (17)

where

$$A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}, \quad I(t) = \sin \frac{2\pi}{T_i} \begin{pmatrix} b_{1,l} \\ b_{2,l} \end{pmatrix}.$$  \hspace{1cm} \lambda < 0.

The corresponding mapping at period $T_i$ can be easily obtained:

$$x_{\tau+1} = A_l x_\tau + B_l,$$  \hspace{1cm} (18)

where

$$A_l = \begin{pmatrix} e^{i\Omega_l} & 0 \\ 0 & e^{i\Omega_l} \end{pmatrix}, \quad B_l = \frac{2\pi T_i}{\lambda T_i^2 + 4\pi^2 (e^{i\Omega_l} - 1)} \begin{pmatrix} b_{1,l} \\ b_{2,l} \end{pmatrix}.$$  \hspace{1cm} (19)

The Lipschitz constant is obtained as follows:

$$\ln p(g_l) = e^{i\Omega_l} < 1.$$  \hspace{1cm} (20)

This shows that the map is a contraction. Therefore a fractal is constructed on the Poincaré section. The similarity dimension, one of fractal dimension, is obtained as follows:

$$d = \frac{-\ln 3}{\lambda T_i}.$$  \hspace{1cm} (20)

This shows that the dimension is proportional to $1/\lambda T_i$ while it is independent of input amplitude $b$. Figure 1 is the $\lambda$-dependence of the dimension $d$. The solid curve is the dimension calculated by Eq. (20), which shows good agreement with the correlation dimension obtained numerically using log-log plot in the totally disconnected region.

![Figure 1: $\lambda$-dependence of the dimension $d$.](image-url)
In the overlapping region, however, the solid curve gradually deviated from the correlation dimension. This is because the dimension of the Poincaré section is the ceiling of 2. The above figures denoted (a), (b), (c), (d), and (e) are examples obtained at the Poincaré section. (a) and (b) are totally disconnected, i.e., clusters of the same color are completely separated from each other. (d) and (e) are overlapping, i.e., the clusters partially overlap each other. (c) is just-touching, i.e., the clusters touch each other at a single point. Although this is a simple example using a linear equation, a nonlinear equation also shows similar characteristics of fractals [10]. Recently we derived multi-fractal formalism under some condition. The result is as follows:

\[ D_q = \frac{n}{q-1} \ln \frac{\sum_{i=1}^{N} p_i^q}{\Delta T}, \]  

where \( D_q \), \( n \), \( N \), \( \lambda \), \( T \), and \( p_i \) are general dimension of \( q \), number of states, number of inputs, statistical contraction, time length of inputs, and probability of \( i \)-th input, respectively.

6. Discussion

In this paper a theoretical framework is reviewed for the continuous dynamical systems stochastically excited by external temporal inputs. In this section we discuss some related works. More general theory has been presented in order to model complex systems that interact strongly with other systems. It has been revealed that these dynamics are generally characterized by fractals when the iterated functions are not the contractions [10]. The hierarchical structure of fractals and the noise effect of inputs have been investigated [7]. The fractals generated by switching vector fields have been observed in different domains such as a forced damped oscillator [11], an electronic circuit [12], artificial neural networks [13], and human behavior [14]. Closure of the fractals in both linear [15] and non-linear systems [16] has been also presented. A set of attractors obtained by periodic inputs can approximate trajectories of fractals [17]. These works show that fractals are indispensable for understanding of dynamics observed in the Hybrid Dynamical Systems as a complex system.

References


Interaction of Fast-Scale and Slow-Scale Bifurcations in Current-Mode Controlled DC/DC Converters

Yanfeng Chen∗†, Chi K. Tse∗ and Shui-Sheng Qiu†

Abstract—This paper investigates the interaction of fast-scale and slow-scale bifurcations in the boost converter under current-mode control operating in continuous conduction mode. Effects of varying some chosen parameters on the qualitative behaviors of the system are studied in detail. Boundaries of stable region, slow-scale bifurcation region, fast-scale bifurcations region, interacting fast and slow-scale bifurcation region are identified.

1. Introduction

The current-mode control scheme is widely used in power converters [1]. Bifurcation behaviors in dc/dc converters under current-mode control have been reported recently [2, 3]. Generally, two distinct types of bifurcations have been identified for such circuits, namely slow-scale bifurcation and fast-scale bifurcation. The slow-scale bifurcation can be regarded as a kind of low-frequency oscillation which is caused by the voltage feedback loop permitting low-frequency oscillation [4, 5, 6]. The fast-scale bifurcation, which is caused by inner current loop instability, is often found in current-mode controlled converters, and it manifests as period-doubling in the time scale, as reported in Iu et al. [7] for parallel boost converters, Wong et al. [8] for noise-coupled boost converters, and Wu et al. [9] for power-factor-correction converters. The fast-scale and slow-scale bifurcations have been independently investigated; however, their interactions are not thoroughly studied despite the frequent occurrence of such phenomena.

Fundamentally the two mechanisms responsible for the fast-scale and slow-scale bifurcations can actually interact each other when the parameters chosen fall in the right region. Our objective in this paper is to study such interactions and to identify the boundaries of stable region, slow-scale bifurcation region, fast-scale bifurcation region, and the interacting region in terms of some chosen parameters.

2. Current-Mode Controlled Boost Converter

The current-mode controlled boost converter is shown in Fig. 1(a). The system has an outer voltage loop and an inner current loop. The voltage loop, which consists of an error amplifier (EA), a compensation network, provides the reference for the inner current loop. The inner current-loop consists of a current transformer, a compensation ramp signal and a current sense amplifier. The compensation ramp is added to stabilize the converter if a wide range of output voltage is required [6]. The output of the two loops is then connected to the inputs of the comparator whose output is used to reset a flip-flop latch to give a pulse-width modulated waveform to control switch $S_T$. The operation can be briefly described as follows. The flip-flop latch is set periodically by the clock signal, turning on the switch $S_T$. Then, the inductor current goes up linearly, and is compared with a reference level, which is equal to the output of the error amplifier of the voltage loop minus the compensation ramp signal. When the peak inductor current reaches the reference level, the output of the comparator resets the flip-flop, thereby turning off the switch. When the switch is off, the inductor current falls almost linearly if the output capacitor is sufficiently large. The cycle repeats when the flip-flop is set again by the clock. Typical waveforms of are shown in Fig. 1(b), where $m_1$ is the slope of the compensation ramp signal, and $m_1$ and $m_2$ are the rising and falling slopes of the inductor current with the switch $S_T$ on and off respectively.

3. A Glimpse at Bifurcation Behavior

We begin with a series of typical waveforms from exact cycle-by-cycle computer simulations to show some possible bifurcation
beverages in this system. The main parameters affecting fast-scale bifurcations are $L/E$ and $m_c$, whereas those affecting slow-scale bifurcations are the voltage feedback gain and time constant, as studied previously [10]. To facilitate simulation study, we choose parameter values as listed in Table 1.

For a certain range of parameters, normal periodic operation, slow-scale bifurcation, fast-scale bifurcation, coexisting fast and slow-scale bifurcation, and “saturated” operation (border collision) can be observed. Typical waveforms are shown in Figs. 2, 3, and 4. Consistent with usual understanding, we can see from these waveforms that slow-scale bifurcations can be eliminated by increasing the feedback gain or time constant, and that fast-scale bifurcations can be eliminated by increasing $m_f$ or $m_c$.

### Table 1: Circuit parameters for simulation study

<table>
<thead>
<tr>
<th>Component/parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input voltage $E$</td>
<td>3 – 25 V</td>
</tr>
<tr>
<td>Inductance $L$</td>
<td>120 – 195 $\mu$H</td>
</tr>
<tr>
<td>Capacitance $C$</td>
<td>2000 $\mu$F</td>
</tr>
<tr>
<td>Load resistance $R$</td>
<td>3 – 20 $\Omega$</td>
</tr>
<tr>
<td>Switching frequency $f_s$</td>
<td>25 kHz</td>
</tr>
<tr>
<td>Reference output voltage $V_{\text{ref}}$</td>
<td>1.8 V</td>
</tr>
<tr>
<td>Voltage divider $R_1$, $R_2$</td>
<td>47.5 k$\Omega$, 2.5 k$\Omega$</td>
</tr>
<tr>
<td>Compensation network $R_m$, $C_a$</td>
<td>72.3 k$\Omega$, 0.23 $\mu$F</td>
</tr>
<tr>
<td>Compensation ramp $V_p$</td>
<td>0.25 V</td>
</tr>
<tr>
<td>Inductance current sampling gain $M$</td>
<td>0.082 V/A</td>
</tr>
</tbody>
</table>

### 4. Detailed Examination: Boundaries of Operations

In this section, we take a detailed look into the qualitative behaviors of the system, and present the boundaries of stable region, slow-scale bifurcation region, fast-scale bifurcation region, and coexisting (interacting) fast and slow-scale bifurcation region in terms of selected circuit parameters.

#### Effects of Varying Input Voltage $E$ and Load Resistance $R$

The behavior of the system varies with $E$ and $R$. Two perspectives of operating boundaries are shown in Fig. 5(a), where the boundaries divide regions of stable operation and “saturated” region. The transition from the stable region to the other is a Hopf type bifurcation. Parameters except $R$ and $E$ are kept as listed in Table 1.

#### Effects of Varying Rising Slope of Inductor Current $E/L$

The rising slope of the inductor current, $E/L$, has significant influence on the operation of the system. For each pair of $E$ and $R$ used in simulation, we choose those points located at the boundary curve in Fig. 5(a), and change $L$ while keeping all other parameters as listed in Table 1. The simulated operation boundaries are shown in Fig. 5(b) in two perspectives.

#### Effects of Varying Feedback Gain $g$ and Time Constant $\tau_f$

The feedback gain and time constant have obvious influence on the stability, especially for the slow-scale bifurcation because of the bandwidth limitation imposed by the time constant. We keep
E and R constant, and vary L/E. We can then change g and τf by varying Rs and Cs. Figure 6 shows the stability boundaries for different values of L/E, corresponding to the slow-scale bifurcation boundary, stable region and fast-scale bifurcation region of Fig. 5(a). In Fig. 6, the regions under the boundary curves correspond to the “saturated” operation, and the operations above the boundaries depend on the values of L/E and g. As seen from Fig. 6, interacting slow and fast-scale bifurcations can be observed for some parameter ranges corresponding to relatively small L/E.

In a likewise manner, another set of boundaries can be collected for fixed L. As limited by space, they are omitted here.

**Effects of Varying Compensation Ramp mc**

It has been known that the compensation ramp strongly affects the instability. Varying mc will lead to interacting slow and fast-scale bifurcations. In our simulation, we vary mc by changing the ramp amplitude Vr. Boundary of operations are plotted in Fig. 7, from which we see clearly that both fast-scale and interacting fast and slow-scale bifurcations are possible. It is worth noting that varying mc has little effect on slow-scale bifurcation, such as the period of the limit cycle and the amplitude of the low-frequency oscillations, as shown in Fig. 4.

5. Conclusion

In this paper we have studied the interaction of slow and fast-scale bifurcations. Basically we have observed that under certain parameter ranges, current-mode controlled boost converters can be fast-scale and slow-scale unstable simultaneously. Such interacting bifurcation occurs when the low-frequency voltage loop and the fast inner current loop become unstable simultaneously. In this paper we have identified the parameter ranges under which slow-scale, fast-scale, and interacting slow and fast-scale bifurcations occur.

**Acknowledgment**

This work is supported by Hong Kong Polytechnic University under a Research Grant for Chair Professors (No. 1-BBZA).

**References**


Figure 5: Operating boundaries with (a) varying $E/R$ and $D$; (b) varying $L/E$.

Figure 6: Operating boundaries with varying feedback gain and time constant. $E = 6.1877\, \text{V}$, $R = 10.78\, \Omega$.

Figure 7: Operating boundaries with varying compensation slope and $\tau_a = C_a R_a = g\tau_f$.


Averaging Method Analysis for the Nonlinearly-Coupled van der Pol and Rayleigh Oscillators

Toshihiro Hishida†, Kuniyasu Shimizu‡ and Tetsuro Endo†

†Department of Electronics and Communication, Meiji University
1-1-1 Higashi-mita, Tama-ku, Kawasaki-shi 214-8571, Japan
Email: endoh@isc.meiji.ac.jp

Abstract — In this paper, we investigate the synchronization characteristics of a coupled system of a van der Pol oscillator and a Rayleigh oscillator. In particular, we assume the nonlinear coupling scheme proposed by S. Nagano. This coupling scheme was invented from an aggregation of slime, kind of ameba, and is known to show strong synchronization ability. Assuming that the nonlinearity is weak, we analyze such a system by using averaging method and clarify the mechanism for strong synchronization.

1. Introduction

Coupled oscillators have been investigated for a long time by many researchers [1]. Almost all of them assumed linear coupling scheme which was a natural configuration in view of engineering applications including electronics circuits and mechanical systems [2]. Recently, S. Nagano proposed a nonlinear coupling scheme between oscillators from the viewpoint of biological system such as an aggregation of slime, a kind of ameba, [3,4]. One of the characteristics of this nonlinear coupling scheme is its strong synchronization ability. Namely, different kind of oscillators with different natural frequencies can be synchronized easily. For example, we will investigate a coupled system of van der Pol and Rayleigh oscillators. Unfortunately, only computer simulation results were presented for this example. Therefore, by assuming weak nonlinearity, we will analyze this system with different natural frequencies via averaging method [5] to elucidate the reason of this strong synchronization ability.

2. Derivation of the Averaged Equation

In this section we will derive the differential equation describing this system. A van der Pol oscillator with natural angular frequency \( \omega_1 \) can be written as

\[
\dot{x}_1 = y_1 \\ y_1 = -\omega_1^2 x_1 + \varepsilon (1 - x_1^2) y_1 \equiv Y_1(x_1, y_1)
\]

A Rayleigh oscillator with natural angular frequency \( \omega_2 \) can be written as

\[
\dot{x}_2 = y_2 \\ y_2 = -\omega_2^2 x_2 + \varepsilon (1 - \frac{1}{3} x_2^2) y_2 \equiv Y_2(x_2, y_2)
\]

The nonlinearly coupled system can be written as

\[
\dot{x}_1 = X_1(x_1, y_1) \\ y_1 = Y_1(x_1 + \gamma_1 x_1 + x_2, y_1) \\ \dot{x}_2 = X_2(x_2, y_2) \\ y_2 = Y_2(x_2 + \gamma_2 x_1 + x_2, y_2)
\] (3)

where \( \gamma_1 \) and \( \gamma_2 \) denote coupling factors of two oscillators.

Substituting (1) and (2) into (3) yields:

\[
\dot{x}_1 = y_1 \\ \dot{y}_1 = -\omega_1^2 (1 + \gamma_1) x_1 + \gamma_1 x_2 + \varepsilon [1 - (1 + \gamma_1) x_1 + \gamma_1 x_2^2] y_1 \\
\dot{x}_2 = y_2 \\ \dot{y}_2 = -\omega_2^2 \gamma_2 x_1 + (1 + \gamma_2) x_2 + \varepsilon (1 - \frac{1}{3} y_2^2) y_2
\]

Deleting \( y_1 \) and \( y_2 \) in (4) yields:

\[
\dot{x}_1 + \omega_1^2 ((1 + \gamma_1) x_1 + \omega_1^2 \gamma_1 x_2 \\
= \varepsilon [1 - (1 + \gamma_1) x_1 + \gamma_1 x_2^2] \dot{x}_1 \\
\dot{x}_2 + \omega_2^2 ((1 + \gamma_2) x_2 + \omega_2^2 \gamma_2 x_1 + \varepsilon (1 - \frac{1}{3} y_2^2) \dot{x}_2
\]

Equation (5) can be written in the following vector form.

\[
\dot{x} + B x = \varepsilon x - \varepsilon Q G(x, \dot{x})
\]

where

\[
x = [x_1, x_2]^T, \quad B = \begin{bmatrix}
(1 + \gamma_1) \omega_1^2 & \gamma_1 \omega_1^2 \\
\gamma_1 \omega_2^2 & (1 + \gamma_2) \omega_2^2
\end{bmatrix}, \quad G(x, \dot{x}) = [g_1(x, \dot{x}), g_2(x, \dot{x})]^T
\]

\[
g_1(x, \dot{x}) = [(1 + \gamma_1) x_1 + \gamma_1 x_2] \dot{x}_1, \quad g_2(x, \dot{x}) = \frac{1}{3} \dot{x}_2^3
\]

Applying a linear transformation \( x = Py \) to (6) and multiplying \( P^{-1} \) from the left hand side yields:

\[
\dot{y} + (P^{-1} B P) y = \varepsilon \dot{y} - \varepsilon Q G(x, \dot{x}), \quad P \equiv P^{-1}
\] (7)

\[
P = \begin{bmatrix}
p_{11} & p_{12} \\
p_{21} & p_{22}
\end{bmatrix}, \quad Q = \begin{bmatrix}
q_{11} & q_{12} \\
q_{21} & q_{22}
\end{bmatrix}
\]

By choosing an appropriate \( P \), Equation (7) can be diagonalized as

\[
\dot{y} + \begin{bmatrix}
\Omega_1^2 & 0 \\
0 & \Omega_2^2
\end{bmatrix} y = \varepsilon \dot{y} - \varepsilon Q G(x, \dot{x}), \quad y = [y_1, y_2]^T
\] (8)
where $\Omega_1^2$ and $\Omega_2^2 (> \Omega_1^2)$ are eigenvalues of $B$.

Equation (8) can be written in the following scalar form
\[ \ddot{y}_k + \Omega_k^2 y_k = \varepsilon \dot{y}_k - \varepsilon h_k(y, \dot{y}) \quad \text{for} \quad k = 1, 2 \]

(9)

From (6) $h_1$ and $h_2$ can be calculated as
\[ h_1(y, \dot{y}) = q_{11} g_1(x, \dot{x}) + q_{12} g_2(x, \dot{x}) \]
\[ = q_{11}((1 + \gamma_1)x_1 + \gamma_1 x_2\dot{x}_1 + \frac{1}{3} q_{12} \dot{x}_1^3) \]
\[ = q_{11}((1 + \gamma_1)(p_{11}x_1 + p_{12}y_1) + \gamma_1(p_{21}x_1 + p_{22}y_2))^2(p_{11}\dot{y}_1 + p_{12}\dot{y}_2) \]
\[ + \frac{1}{3} q_{12}(p_{21}\dot{y}_1 + p_{22}\dot{y}_2)^3 \]
\[ h_2(y, \dot{y}) = q_{21} g_1(x, \dot{x}) + q_{22} g_2(x, \dot{x}) \]
\[ = q_{21}((1 + \gamma_1)x_1 + \gamma_1 x_2\dot{x}_1 + \frac{1}{3} q_{22} \dot{x}_1^3) \]
\[ = q_{21}((1 + \gamma_1)(p_{11}x_1 + p_{12}y_1) + \gamma_1(p_{21}x_1 + p_{22}y_2))^2(p_{11}\dot{y}_1 + p_{12}\dot{y}_2) \]
\[ + \frac{1}{3} q_{22}(p_{21}\dot{y}_1 + p_{22}\dot{y}_2)^3 \]

By assuming $\varepsilon = 0$ in (9), $y_k$ and $\dot{y}_k$ can be derived easily as
\[ y_k = \rho_k \sin(\Omega_k t + \theta_k), \quad \dot{y}_k = \rho_k \Omega_k \cos(\Omega_k t + \theta_k) \quad \text{for} \quad k = 1, 2 \]

(12)

When $\varepsilon$ is non-zero and small, the amplitude $\rho_k$ and the phase $\theta_k$ can be replaced as slowly-varying functions of time such as $\rho_k(t)$ and $\theta_k(t)$, respectively. From averaging theory, averaged equations become as follows
\[ \dot{\rho}_k = \frac{\varepsilon}{\Omega_k} (f_k(y, \dot{y}) \cos(\Omega_k t + \theta_k)) \]
\[ \dot{\theta}_k = -\frac{\varepsilon}{\Omega_k \rho_k} \]

(13)

where $\langle \cdot \rangle$ denotes the time average from 0 to infinity.

By using (9), equation (13) can be calculated as
\[ \dot{\rho}_1 = \frac{1}{2} \varepsilon p_1(1 - \frac{1}{4} A p_1^2 - \frac{1}{6} B p_2^2) \equiv F_1(p_1, p_2) \]
\[ \dot{\rho}_2 = \frac{1}{2} \varepsilon p_2(1 - \frac{1}{2} C p_1^2 - \frac{1}{4} D p_2^2) \equiv F_2(p_1, p_2) \]

(14)

where
\[ A \equiv [(1 + \gamma_1)p_{11} + \gamma_1 p_{21}]^2 p_{11}q_{11} + p_{21}^2 q_{12}\Omega_1^2 \]
\[ B \equiv [(1 + \gamma_1)p_{12} + \gamma_1 p_{22}]^2 p_{11}q_{11} + p_{22}^2 q_{12}\Omega_2^2 \]
\[ C \equiv [(1 + \gamma_1)p_{11} + \gamma_1 p_{21}]^2 p_{12}q_{21} + p_{21}^2 p_{22}q_{12}\Omega_1^2 \]
\[ D \equiv [(1 + \gamma_1)p_{12} + \gamma_1 p_{22}]^2 p_{12}q_{21} + p_{22}^2 q_{22}\Omega_2^2 \]

(15)

In deriving (14), we assume the nonresonant condition between $\Omega_1$ and $\Omega_2$ such as
\[ 3\Omega_1 \neq \Omega_2. \]

3. Stability of the Same-Phase and the Reverse-Phase Solutions

From the averaged equation (14), the stability of the same-phase solution can be obtained by assuming $\rho_1 = \rho_2 = 0$ and $\rho_1 = 0$ and $\rho_2 \neq 0$. This gives a steady-state solution as follows
\[ \rho_01 = 0, \quad \rho_02 = 4 - \frac{4}{\mu} \]

(16)

The stability of (16) can be judged from the Jacobian matrix $J$
\[ J = \begin{bmatrix} \frac{\partial F_1}{\partial p_1} & \frac{\partial F_1}{\partial p_2} \\ \frac{\partial F_2}{\partial p_1} & \frac{\partial F_2}{\partial p_2} \end{bmatrix} \]

(17)

where $F_1$ and $F_2$ are given in (14).

Namely, if all eigenvalues of $J$ have negative real parts, the solution is asymptotically stable, and if one of the eigenvalue of $J$ have a positive real part, it is unstable. In the same manner, the stability of the reverse-phase solution can be obtained by assuming $\rho_1 \neq 0$ and $\rho_2 = 0$; namely,
\[ \rho_01 = 4 - \frac{4}{\mu}, \quad \rho_02 = 0 \]

(18)

Actual procedure for obtaining the same-phase and the reverse-phase solution with their stability judgement is as follows. 1) Determine all parameters $\gamma_1$, $\gamma_2$, $\omega_1$, and $\omega_2$. 2) Calculate eigenvalues $\Omega_1^2$ and $\Omega_2^2 (> \Omega_1^2)$ of $B$ numerically. 3) Determine $P$ and $Q \equiv P^{-1}$ from eigenvectors associated with $\Omega_1^2$ and $\Omega_2^2$. 4) Obtain the same-phase and reverse-phase solutions from (16) and (18) with their stability judgement. Henceforth, we define $\gamma \equiv \gamma_1 = \gamma_2$.

Figure 1 presents the synchronization characteristics of the same-phase solution $\rho_02$ in terms of $\log_{10}(\omega_2/\omega_1)$ for various values of $\gamma$ with $\omega_1 = 1$. Note that the synchronization characteristics are asymmetric with respect to the origin. This is in contrast to the case of two nonlinearly coupled van der Pol oscillators, where the synchronization characteristics are symmetric with respect to the origin.

Figure 1: Synchronization characteristics $\rho_02$ versus $\log_{10}(\omega_2/\omega_1)$ for various values of coupling factor $\gamma$ for the same-phase solution
Amplitude of $x_1$ versus log$_{10}(\omega_2/\omega_1)$ for various values of $\gamma$ for the same-phase solution

Figure 2: Amplitude variation $x_1$ and $x_2$ in terms of log$_{10}(\omega_2/\omega_1)$ for various values of $\gamma$ for the same-phase solution

Synchronized frequency $\Omega_2$ in terms of log$_{10}(\omega_2/\omega_1)$ for various values of $\gamma$ for the same-phase solution

Figure 3: Synchronized frequency $\Omega_2$ in terms of log$_{10}(\omega_2/\omega_1)$ for various values of $\gamma$ for the same-phase solution

Table 1: Comparison between computer simulation results and averaging method results for the same-phase solution

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\Omega_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer simulation</td>
<td>1.088</td>
<td>1.235</td>
<td>1.434</td>
</tr>
<tr>
<td>Averaging method</td>
<td>1.091</td>
<td>1.228</td>
<td>1.436</td>
</tr>
</tbody>
</table>

(a) $\gamma = 0.5$, $\omega_2 = 1.03$, $\omega_1 = 1.0$, $\varepsilon = 0.1$

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$\Omega_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer simulation</td>
<td>0.319</td>
<td>0.944</td>
<td>2.219</td>
</tr>
<tr>
<td>Averaging method</td>
<td>0.322</td>
<td>0.942</td>
<td>2.219</td>
</tr>
</tbody>
</table>

(b) $\gamma = 1.0$, $\omega_2 = 1.45$, $\omega_1 = 1.0$, $\varepsilon = 0.5$

Figures 2 (a) and (b) show the amplitude characteristics of $x_1$ and $x_2$ with respect to log$_{10}(\omega_2/\omega_1)$ with $\omega_1 = 1$. When $\omega_2 < \omega_1$, the amplitude of $x_1$ is larger than that of $x_2$, and when $\omega_2 > \omega_1$ the amplitude of $x_2$ is larger than that of $x_1$.

Figure 3 shows the synchronized angular frequency $\Omega_2$ of the same-phase solution with respect to log$_{10}(\omega_2/\omega_1)$. Notice that $\Omega_2$ is much larger than $\omega_1$ and $\omega_2$ especially for large $\gamma$.

Figures 4 (a) and (b) show computer simulation results of the same-phase solution for various $\gamma$ and $\omega_2$ with $\omega_1 = 1$ and $\varepsilon = 0.1$ being fixed. They all show synchronization with the same-phase in steady state. Table 1 compares the result of averaging method with that of computer simulation. They show in good agreement.

Figure 5 shows the synchronization characteristics of the reverse-phase solution in terms of $\rho_{01}$ versus log$_{10}(\omega_2/\omega_1)$ for various values of $\gamma$ with $\omega_1 = 1$.

Figure 6 shows the synchronized angular frequency $\Omega_1$ of the reverse-phase solution with respect to log$_{10}(\omega_2/\omega_1)$. Figure 7 shows the computer simulation results of the reverse-phase solution for a certain $\gamma$ and $\omega_2$. They agree well with the results of averaging method as seen in Table 2.
4. Conclusions

We investigate the synchronization characteristics of the coupled system of a van der Pol and Rayleigh oscillators. We use the averaging method for analyzing this system. We confirm by direct computer simulation for $\epsilon = 0.1$ and $0.5$ that averaging method gives accurate result. It should be noted that the synchronization range of this system is smaller than that of two mutually coupled van der Pol oscillators and that the synchronization range is asymmetric with respect to the origin. So far, we do not know whether or not this coupled systems has a practical meaning in view of biology. However, this system is a good example demonstrating strong synchronization ability. For future problem, we will investigate several number of coupled oscillators in which van der Pol and Rayleigh oscillators are mixed.

References


Averaging Method Analysis for Three-Coupled Van Der Pol Oscillators with Nonlinear Coupling

Kuniyasu Shimizu†, Tetsuro Endo‡ and Hisa-Aki Tanaka‡

†Department of Electronics and Communication, Meiji University
1-1-1 Higashi-Mita, Tama-ku, Kawasaki-city 214-8571 JAPAN
‡Department of Electronic Engineering, The University of Electro-Communications(UEC)
1-5-1 Choufu-gaoka, Choufu-city, Tokyo 182-8585 JAPAN
Email: shimizu[endo@isc.meiji.ac.jp, htan@ee.uec.ac.jp

Abstract—In this paper, we investigate synchronization characteristics of three nonlinearly-coupled van der Pol oscillators. This coupling originates in an aggregation of slime (a kind of ameba), and presents strong synchronization ability against detuning. We clarify that an arbitrary number of oscillators coupled in this manner can be reduced to comparatively simple averaged equation. For example, we demonstrate the synchronization range in terms of detuning for three-coupled van der Pol oscillators. As a result, we confirm large synchronization range in case of \( \omega_1 = \omega_2 = 1 \) and \( \omega_3 = 0.01 - 100 \) and in case of \( \omega_1 = 1 , \omega_2 = 10 \) and \( \omega_3 = 0.01 - 100 \) for example.

1. Introduction

Synchronization of rhythms plays a central role in system control both in science and technology. Recently, mutual synchronization scheme based on the uni-cellular slime ameba is proposed in the field of biology[1,2]. The uni-cellular slime ameba is essentially a single-organism under starvation. In the simplest case, we assume this ameba as a limit cycle oscillator cell \( j \) such as

\[
\frac{dP_j}{dt} = F_j(P_j, R_j), \quad \frac{dR_j}{dt} = G_j(P_j, R_j)
\]

(1)

where \( P_j \) denotes the product density and \( R_j \) denotes activity of receptor. The aggregated living organism consisting of \( N \)-cells becomes a coupled system as follows

\[
\begin{align*}
\frac{dP_j}{dt} &= F_j(P_j, R_j) \\
\frac{dR_j}{dt} &= G_j(P_j + \gamma \sum_{i=1}^{N} P_i, R_j) \quad j=1,2,\ldots,N
\end{align*}
\]

(2)

where \( \gamma \) denotes a coupling factor. This becomes nonlinear coupling generally. In [2], powerful synchronization ability of this nonlinear coupling scheme has been demonstrated for various mutually coupled oscillators via computer simulation.

Mutual synchronization of oscillators has been investigated for a long time via various methods. Most of synchronization problems assume, however, linear coupling between oscillators [3,4]. In our previous work, we showed that the averaging method could be applied in the case of two mutually-coupled van der Pol oscillators with this nonlinear coupling scheme, and showed the synchronization range could be obtained theoretically for weakly nonlinear case via averaging method [5]. In this work, we obtain averaged equation for an arbitrary number of oscillators. For example, we demonstrate the synchronization characteristics for three nonlinearly-coupled van der Pol oscillators by investigating the stability of various equilibrium points of the averaged equation.

2. Analysis via Averaging Method

From [5], \( N \)-mutually coupled van der Pol oscillators can be written as follows:

\[
\ddot{x} + Bx = \varepsilon \dot{x} - \varepsilon G(x, \dot{x})
\]

(3a)

where

\[
x = [x_1, x_2, \ldots, x_N]^T
\]

\[
B = \begin{pmatrix}
(1 + \gamma)\omega_1^2 & \gamma \omega_1^2 & \cdots & \gamma \omega_1^2 \\
\gamma \omega_2^2 & (1 + \gamma)\omega_2^2 & \cdots & \gamma \omega_2^2 \\
\vdots & \vdots & \ddots & \vdots \\
\gamma \omega_N^2 & \gamma \omega_N^2 & \cdots & (1 + \gamma)\omega_N^2
\end{pmatrix}
\]

\[
G(x, \dot{x}) = [g_1(x, \dot{x}), g_2(x, \dot{x}), \ldots, g_N(x, \dot{x})]^T
\]

(3b)

and where

\[
g_k(x, \dot{x}) = \left(x_k + \gamma \sum_{i=1}^{N} x_i \right)^2 \dot{x}_k
\]

(3c)

We define the eigenvalues of \( B \) as \( \Omega_1^2 < \Omega_2^2 < \cdots < \Omega_N^2 \) and the associated eigenvectors as \( p_j = [p_{j1}, p_{j2}, \cdots, p_{jN}]^T \) for \( j = 1, 2, \ldots, N \), respectively. Applying a non-singular linear transformation \( x = Py \) to (3a) and multiplying \( P^{-1} \) from the left-hand side gives the following transformed equation.

\[
\ddot{y} + (P^{-1}BP)y = \varepsilon \dot{y} - \varepsilon P^{-1}G(y, \dot{y})
\]

(4a)
\[ PBP^{-1} = \begin{pmatrix}
\Omega_2 & 0 \\
0 & \ddots \\
& \ddots & \ddots \\
& & 0 & \Omega_2
\end{pmatrix} \]
\[ P = \begin{pmatrix}
p_{11} & \cdots & p_{1N} \\
\vdots & \ddots & \vdots \\
p_{N1} & \cdots & p_{NN}
\end{pmatrix} \]

Equation (4a) can be written in the following scalar form:
\[ \ddot{y}_k + \Omega_k^2 y_k = e\dot{y}_k - e f_k(y, \dot{y}) \quad (5a) \]
where functions \( f_k \) are expressed in terms of \( y, \dot{y} \) through the transformation \( x = Py \) as follows:
\[ h_k (y, \dot{y}) = \sum_{m=1}^{N} q_{km} g_k(x, \dot{x}) = \frac{e}{\partial t} \sum_{m=1}^{N} (p_{mn} + \gamma \sum_{m=1}^{N} p_{mn} y_m) \]
\[ = \sum_{m=1}^{N} \sum_{n=1}^{N} b(k,m_1,m_2,m_3) y_{m_1} y_{m_2} y_{m_3} \]
where \( b(k,m_1,m_2,m_3) = \sum_{m=1}^{N} q_{km} g_k(x, \dot{x}) a_{m_2} p_{mn} \)

In averaging method, we first assume \( \epsilon = 0 \) in (5a), and obtain the unperturbed solution. It can be calculated easily as \( \dot{y}_k = \rho_k \sin(\Omega_k t + \theta_k), y_k = \rho_k \cos(\Omega_k t + \theta_k) \) for \( k = 1, 2, \ldots, N \) [6].

When \( \epsilon \neq 0 \), we assume \( \rho_k \) and \( \theta_k \) as functions of time. From the theory of averaging, the dynamics of these functions can be calculated from the following averaged equation for \( k = 1, 2, \ldots, N \) [6]:
\[ \rho_k = \frac{1}{\epsilon} \left( f_k(y, \dot{y}) \cos(\Omega_k t + \theta_k) \right) \]
\[ \theta_k = \frac{1}{\epsilon} \left( f_k(y, \dot{y}) \sin(\Omega_k t + \theta_k) \right) \]

where \( \langle \cdot \rangle \) denotes time average from zero to infinity.

Equation (6) can be calculated by assuming for the non-resonant case for \( k = 1, 2, \ldots, 3 \) as follows:
\[ \dot{\rho}_k = \frac{1}{\epsilon} \left( 1 - \sum_{m=1}^{N} b(k,m_1,m_2,m_3) y_{m_1} y_{m_2} y_{m_3} \right) \]
\[ \dot{\theta}_k = 0. \]

Since \( \theta_k \) can be determined as an arbitrary constant, we only investigate the amplitude equation. Therefore, by defining \( U_k \equiv \rho_k^2 \), (7) can be simplified as
\[ U_k = \epsilon F_k \left( \sum_{m=1}^{N} b(k,m_1,m_2,m_3) U_{m_1} y_{m_2} y_{m_3} \right) \]
\[ \equiv \epsilon F_k(U_1, U_2, \cdots, U_N) \]

for \( k = 1, 2, \ldots, N \).

The non-resonant condition can be written as follows for \( m_1, m_2, m_3 = 1, 2, \cdots, N \).

(1) \( \Omega_{m_1} \pm \Omega_{m_2} \pm \Omega_{m_3} \pm \Omega_0 = 0 \) where \( m_1 \neq m_2 \neq m_3 \neq k \)
(2) \( 2\Omega_k \pm \Omega_{m_1} \pm \Omega_{m_2} = 0 \) where \( m_1 \neq m_2 \neq k \)
(3) \( 3\Omega_k - \Omega_{m_1} = 0 \) where \( k \neq m_3 \)

3. Analysis of Steady-State Solutions via Averaging Method for Three-Coupled Oscillator Case

The steady-state solution corresponds to an equilibrium point in (8). For example, assuming three mutually-coupled van der Pol oscillators, there are seven equilibrium points \((U_{01}, U_{02})\) obtained by assuming \( U_k = 0 \) for \( k = 1, 2, \ldots, N \) in (8). Namely, \((a)U_1 = U_2 = U_3 = 0, (b)U_1 \neq 0, U_2 = U_3 = 0, (c)U_2 \neq 0, U_1 = U_3 = 0, (d)U_3 \neq 0, U_1 = U_2 = 0, (e)U_1 \neq 0, U_2 \neq 0, U_3 = 0, (f)U_1 \neq 0, U_2 = 0, U_1 \neq 0 \) and \((g)U_1 = 0, U_2 \neq 0, U_3 \neq 0\).

From the results of numerical calculation, point (d) corresponds to the same-phase solution. In this paper, we will concentrate to the same-phase solution. This is because the actual coupled system of uni-cellular slime ameba almost presents the same-phase synchronization. The stability of each equilibrium point can be obtained by using the following Jacobian matrix \( J \):

\[ J = \begin{pmatrix}
\frac{\partial F_1}{\partial U_1} & \cdots & \frac{\partial F_1}{\partial U_N} \\
\vdots & \ddots & \vdots \\
\frac{\partial F_N}{\partial U_1} & \cdots & \frac{\partial F_N}{\partial U_N}
\end{pmatrix} \]

Namely, if all eigenvalues of \( J \) have negative real parts, the solution is asymptotically stable. If one of them has positive real part, it is unstable. In this manner, we can judge the stability of each equilibrium point.

We first determine parameters \( \gamma, \omega_1, \omega_2 \) and \( \omega_3 \) in original equation (3a). By fixing all parameters, matrix \( B \) can be determined, and then one can calculate the eigenvalues \( \Omega_1^2, \Omega_2^2 \) and \( \Omega_3^2 \) with matrix \( P \) and \( P^{-1} \) through numerical calculation. Then all necessary parameters in (5b) can be obtained in order. The same-phase solution takes the following form:

\[ U_3 \neq 0, U_1 = U_2 = 0 \Rightarrow U_{03} = \frac{4}{b(3,3,3)}, U_{01} = U_{02} = 0 \]
\[ y_3 = \sqrt{\frac{4}{b(3,3,3)}} \sin(\Omega_3 t + \theta_3), y_1 = y_2 = 0 \]
\[ x_1 = p_{13} \sqrt{\frac{4}{b(3,3,3)}} \sin(\Omega_3 t + \theta_3), x_2 = p_{23} \sqrt{\frac{4}{b(3,3,3)}} \sin(\Omega_3 t + \theta_3) \]
\[ x_3 = \frac{3}{b(3,3,3)} \sin(\Omega_3 t + \theta_3) \]

In (11), the phase \( \theta_3 \) is arbitrary because there is no basin of attraction for this phase since \( \theta_k = 0 \) for \( k = 1, 2, 3 \). Actually, it is determined from the initial condition of (3a).
**4. Numerical Results for The Same-Phase Solution**

We will investigate the synchronization range for the same-phase solution. First, we calculate an example fixing $\omega_1 = \omega_2 = 1.0$ and varying $\omega_3$ from 0.01 to 100.0. Figures 1 (a), (b) and (c) present $\rho_3$ in terms of $\log_{10}(\omega_3/\omega_1)$. Both ends of each curve presents the synchronization limit. In general, for larger coupling factor $\gamma$, synchronization range becomes larger. Different from our previous two-coupled case, these curves are asymmetric with respect to the origin. Namely, the synchronization range below 0 is larger than that above 0. Figures 2 (a), (b) and (c) show the amplitude variation of $x_1$, $x_2$ and $x_3$ in the synchronized state for various values of $\gamma$. From this figure, it is noted that when the ratio $(\omega_3/\omega_1) < 1$, the amplitude of lower frequency of oscillation is suppressed. Figure 3 presents synchronized angular frequency $\Omega_3$ in terms of $\log_{10}(\omega_3/\omega_1)$. In general, $\Omega_3$ is much larger than each individual angular frequency $\omega_1$, $\omega_2$ and $\omega_3$. These results can be confirmed by direct computer simulation of (3a) as shown in Figure 4. The higher frequency of oscillation $(\omega_3)$ suppresses the lower frequency of oscillation $(\omega_1, \omega_2)$ for achieving synchronization. From the view point of dynamics of slime, this may be interpreted as follows. Namely, an oscillator with low natural frequency corresponds to inactive slime cell and an oscillator with high natural frequency corresponds to active slime cell. When they are coupled the contribution of the active cell becomes larger than that of the inactive cell. The amplitudes agree well quantitatively except some parameter points close to the margin of synchronization range where numerical simulation algorithm becomes inaccurate.

Next, we will show another example fixing $\omega_1 = 1.0$,
Figure 3: Synchronized angular frequency variation in terms of $\log_{10}(\frac{\omega_3}{\omega_1})$ fixing $\omega_1=\omega_2=1.0$ for various values of coupling factor $\gamma$.

Figure 4: Computer simulation for synchronization among three van der Pol oscillators for $\varepsilon = 0.1, \gamma = 2.3, \omega_1 = \omega_2 = 1.0$ and $\omega_3 = 8.0$.

$\omega_2 = 10.0$ and varying $\omega_3$ from 0.01 to 100.0. Figures 5(a) and (b) present $\rho_3$ in terms of $\log_{10}(\frac{\omega_3}{\omega_1})$. In this case, synchronization range can not be obtained for comparatively small coupling factor $\gamma$. Synchronization range can be obtained for $\gamma$ greater than 1.2. When $\omega_3$ is near $\omega_2 (\approx 10.0 \gg \omega_1 = 1.0)$, synchronization of each oscillator is easily achieved. But, when $\omega_3$ is near $\omega_1$, they are not synchronized. Different from the first example, the minimum point of the synchronization curve is around $\log_{10}(\frac{\omega_3}{\omega_1}) = 1$, namely $\omega_3 = 10$. This means that synchronization is easier for $\omega_3$ around 10 rather than around 1.

5. Conclusions

We investigate synchronization characteristics of three nonlinearly-coupled van der Pol oscillators with new coupling scheme. We obtain the synchronization range of the same-phase solution against detuning for various coupling factor using averaging method. In the future, we will investigate the synchronization characteristics of a large number of coupled oscillators by using this method. Moreover, we will investigate the synchronization characteristics of strong nonlinear case, and compare it with that for weak nonlinear case.

References

Abstract – General methods for master-slave synchronization and mutual synchronization respectively are presented. The methods can be adopted for the teaching of the topic. Numerical results are given for the synchronization of two Colpitts oscillators.

1. Introduction

Synchronization is a fascinating phenomenon in nature [1, 2, 3] and a useful one in human activity [4]. Recently it was investigated in the modern topic of complex dynamic networks [5, 6, 7, 8, 9, 10, 11]. There are known several types of synchronization. In the following we are concerned only with master-slave synchronization and mutual synchronization of continuous nonlinear dynamical systems. Several theoretical, numerical and experimental results are already reported in literature. However, there are still needed highly elaborated and mathematically based methods [12, 13]. These methods should give precise coupling between systems in order to synchronize. Here we present one method for master-slave synchronization and one for mutual synchronization of identical oscillators. Both methods are general and could be adopted for the teaching of the topic. This is why we call them didactic. We apply these methods to the synchronization (master-slave and mutual respectively) of two Colpitts oscillators.

2. Master-slave Synchronization

Let’s consider a general master system:
\[ \frac{dx}{dt} = F(x), \quad x \in \mathbb{R}^n \quad (1) \]
The slave system is coupled to the master:
\[ \frac{dx}{dt} = F(x) + (H - dF(X)/dX)(x - X) \quad (2) \]
where H is a constant Hurwitz matrix (a matrix with negative real part eigenvalues). The proof is simple. Let’s consider the error \( e = x - X \) and the Taylor expansion:
\[ F(x + e) = F(x) + dF(x)/dx e + \ldots \quad (3) \]
Subtracting (1) from (2) and taking into account (3) we obtain
\[ \frac{dx}{dt} = He \quad (4) \]
We have to emphasize that the matrix H should be chosen in such a manner in order that the coupling to be as simple as possible. When \( dF_i(X)/dX_k \) is constant then it can be chosen \( H_{ik} = dF(X)/dX_k \). When \( dF_i(X)/dX_k \) is variable then \( H_{ik} \) can be chosen \( H_{ik} = p \), where p is a parameter that should be determined in order that the matrix H is a Hurwitz matrix. If the characteristic equation of the matrix H is:
\[ \lambda^3 + a_1 \lambda^2 + a_2 \lambda + a_3 = 0 \]
then the Ruth-Hurwitz conditions are
\[ a_3 > 0, \quad a_2 - a_1 > 0, \quad a_1 > 0. \]
If the above condition can not be fulfilled with one parameter in the matrix H then a second parameter should be introduced in the H matrix. For more details see [12, 15]. Equation (4) assures that \( e = x - X \) converges to zero. We have to note here that the eq. (4) is an approximation because from the Taylor expansion we have taken just one term. If the synchronization is not obtained then new terms should be taken in the Taylor expansion. Such a case will be considered below. Equation (4) is correct and will work for any initial conditions X(0) and x(0) close enough. If the synchronization is not obtained (it means that the initial conditions X(0) and x(0) are not close enough) then new terms should be taken in the Taylor expansion. The present method has been applied to all 19 systems from Sprott collection [14].

3. Mutual Synchronization

Let’s consider two identical general oscillators:
\[ \frac{dx}{dt} = F(x) \quad (5) \]
and
\[ \frac{dy}{dt} = F(y) \quad (6) \]
In order to obtain synchronization it is necessary to couple the two systems. The coupled systems are:
\[ \frac{dx}{dt} = F(x) + u(x,y) \quad (7) \]
\[ \frac{dy}{dt} = F(y) - u(x,y) \quad (8) \]
where
\[ u(x,y) = (H - dF(s)/ds)(x - y)/2 \quad (9) \]
and \( s = (x+y)/2 \) and H a Hurwitz matrix. The proof is simple: We use the notations: \( s = (x+y)/2 \) and \( r = (x-y)/2 \) and the Taylor expansions:
\[ F(s+r) = F(s) + dF(s)/ds r + \ldots \quad (10) \]
\[ F(s-r) = F(s) + dF(s)/ds (-r) \quad (11) \]
Subtracting (8) from (7) we obtain
\[ \frac{dr}{dt} = Hr \quad (12) \]
That assures that r converge to zero and x converge to y it means that the systems synchronize. The present method has been applied [16] to all 19 systems from the Sprott collection [14].

4. Applications to Colpitts Oscillator
4.1 Model of Colpitts Oscillator

The Colpitts Oscillator consists of a bipolar junction, a resonant network with an inductor L and a resistor R and a pair of capacitors $C_1$ and $C_2$. The normalized dynamical model is [17]:

\[
\begin{align*}
\frac{dx_1}{dt} &= \alpha x_2 \\
\frac{dx_2}{dt} &= -\gamma (x_1+x_3)-q x_3 \\
\frac{dx_3}{dt} &= \eta (1+x_2-\exp(-x_1)) \\
\end{align*}
\]

(13)

For $L=100 \mu \text{H}$, $C_1=C_2=47\text{nF}$, $R=45 \Omega$, $I_0=5\text{mA}$ the parameters have the values: $\alpha = 6.2723$, $\gamma = 0.0797$, $q=0.6898$, $\eta = 6.2723$.

4.2 Master-slave Synchronization of two Colpitts Oscillators

We write (1),(2) for the system (13):

\[
\begin{align*}
\frac{dX_1}{dt} &= \alpha X_2 \\
\frac{dX_2}{dt} &= -\gamma (X_1+X_3)-qX_3 \\
\frac{dX_3}{dt} &= \eta (1+X_2-\exp(-X_1)) \\
\end{align*}
\]

(14)

The slave system will be:

\[
\begin{align*}
\frac{dx_1}{dt} &= \alpha x_2 \\
\frac{dx_2}{dt} &= -\gamma (x_1+x_3)-qx_3 \\
\frac{dx_3}{dt} &= \eta (1+x_2-\exp(-x_1)) + (p-\eta \exp(-s))(x_1-x_1)/2 + \eta \exp(-s)(x_1-x_1)^3/6 \\
\end{align*}
\]

(15)

The Hurwitz matrix has $H_{ik} = dF_i/dX_k$ except $H_{31} = p$.

The characteristic equation is:

\[
\lambda^3 + a_1 \lambda^2 + a_2 \lambda + a_3 = 0
\]

with $a_1 = q\alpha = \gamma (\alpha + \eta), a_2 = \eta \gamma, a_3 = p \alpha \gamma$. The Ruth-Hurwitz conditions (see above) give:

\[
0 < p < q(\alpha + \eta)/\alpha
\]

(17)

The coupling in eq. (15) contains 3 terms. Just the first one comes from the presented method. The next 2 terms come from the next terms of the Taylor expansion. This was necessary because the synchronization could not be achieved with the first term.

In Fig. 1 the numerical results are shown for $X_1(0)=X_2(0)=X_3(0)=0.2$ and $x_1(0)=x_2(0)=x_3(0)=-0.2$.

4.3 Mutual Synchronization of Two Colpitts Oscillators

The systems (7), (8) for the system (13) are:

\[
\begin{align*}
\frac{dx_1}{dt} &= \alpha x_2 \\
\frac{dx_2}{dt} &= -\gamma (x_1+x_3)-q x_3 \\
\frac{dx_3}{dt} &= \eta (1+x_2-\exp(-x_1)) + (p-\eta \exp(-s))(x_1-y_1)/2 \\
\end{align*}
\]

(18)

and

\[
\begin{align*}
\frac{dy_1}{dt} &= \alpha y_2 \\
\frac{dy_2}{dt} &= -\gamma (y_1+y_3)-q y_3 \\
\frac{dy_3}{dt} &= \eta (1+y_2-\exp(-y_1)) + (p-\eta \exp(-s))(s-x_1)/2 \\
\end{align*}
\]

(19)

where $s=(x_1+y_1)/2$.

Numerical results are shown in Fig. 2 with $p=0.5, x_1(0)=x_2(0)=x_3(0)=0.2$ and $y_1(0)=y_2(0)=y_3(0)=-0.2$.

Acknowledgements

The authors would like to thank L. Goras, V. Grigoras, A. Maiorescu for helpful discussions.
References

This page has been intentionally left blank.
Transient Analysis for Phase Pattern Switching in Pulse-Driven Star-Coupled Wien-Bridge Oscillators

Seiichiro Moro† and Tadashi Matsumoto‡

†Department of Electrical and Electronics Engineering, University of Fukui
3-9-1 Bunkyo, Fukui 910-8507 JAPAN
‡Department of Electrical and Electronics Engineering, Fukui University of Technology
3-6-1 Gakuen, Fukui 910-8505 JAPAN
Email: moro@ppc8100.fuee.fukui-u.ac.jp, matsu@ccmails.fukui-ut.ac.jp

Abstract—In this paper, the precise transient response in pulse-driven star-coupled Wien-bridge-oscillators. It has reported that there are some difference in the phase pattern switching phenomena between the case when all the oscillators are identical and the case when they are not. From the simulation results using the overlapped limit cycles, the causes of such difference can be clarified.

1. Introduction

There have been many investigations of the synchronization and the multimode oscillation in coupled oscillators [1]–[4]. The synchronization phenomena in resistively coupled oscillators have been reported [2, 3]. In particular, we have investigated the star-coupled system with LC van der Pol oscillators. It also exhibits N-phase oscillations and we can get (N − 1)! different stable phase patterns [3]. Because these “star-coupled” oscillators exhibit a large number of different steady states, they will be used as a structural element of large scale memories and neural networks.

On the other hand, recently, many researches on the networks of the nonlinear elements have been reported [5, 6], because they are important not only as a model for nonlinear systems but also from the viewpoint of biological information processing and their applications. Chua et al. have proposed the network of the sparsely interconnected nonlinear elements called cellular neural network (CNN) [5]. Kaneko has suggested the possibility of the information processing using the network of the nonlinear elements by investigating the globally coupled chaotic maps [6]. Moreover, we have also proposed the coupled oscillators networks for CNNs with star-coupled oscillators [7, 8]. Thus much attention has been paid to the approaches to such applications from the coupled oscillatory systems recently.

When we use the coupled oscillators systems as neural networks and large scale memories, it should be an important problem how to control the systems to get the appropriate phase patterns. To achieve the phase pattern control, we have proposed the star-coupled systems of Wien-bridge oscillators driven by the pulse train and confirmed that the stimulation of the pulse train can cause the phase pattern switching [9, 11]. If the characteristics of all the oscillators are identical, only the phase of the oscillator that is directly connected to the stimulation unit can be switched [9, 10]. On the other hand, if there are some deviations in the circuit parameters in each oscillator, we can see the successive phase pattern switching phenomena in the multiple oscillators [11]. Using these phase pattern switching properties, it will become easier to control those systems. However, the precise mechanisms why those phase pattern switching phenomena can be seen have not been well discussed. In this paper, therefore, we show the precise transient analyses in both systems, and clarify the mechanisms of the phase pattern switching phenomena using the overlapped limit cycles [12]. From the results, we can confirm why only the phase of the single oscillator can be switched in the system in [9], and why the phases of the multiple oscillators can be switched in the system in [11].

2. Circuit Models

The circuit model is shown in Fig. 1. In this circuit, four Wien-bridge oscillators are coupled by one linear resistor r and a stimulation unit which consists of a controlled switch and a DC voltage source is connected to an oscillator. The switch is controlled by the signal sw(t) shown in Fig. 2. While the characteristics of all the subcircuits are identical, we consider the following two cases:

Case 1 $C_1 = C_2 = C_3 = C_4 = C$, so that all the oscillators are identical.

Case 2 $C_k = C + (k - 1)\Delta C$, where $\Delta C$ is the deviation parameter, so that each oscillator has different natural frequencies.

In both cases, it has been reported that the phase pattern switching phenomena can be seen [9, 11]. For case 1, only the phase of the oscillator which the stimulation unit is directly connected can be switched [9], and for the case 2, the successive phase pattern switching phenomena in multiple oscillators can be seen [11]. However, the reasons why the difference between these phenomena in both cases comes from are not precisely discussed. In the following section,
the mechanisms of the phase pattern switching phenomena are shown by using the circuit simulation.

3. Simulation Results

First, we show the simulation results for the Case 1. Note that the stimulation unit is connected to Osc 1. The time waveform is shown in Fig. 3. In this case, the phase pattern has been switched as shown in Fig. 4. In this figure, A, B, and C mean the phases of the oscillators, where A is the phase of the reference signal Osc 4 and B and C indicate 120° delay and forward, respectively.

Now we show more precise motion of the oscillators using the overlapped limit cycle [12]. Obviously, each oscillator has a limit cycle, and its shape should be identical each other. Therefore, we can consider the virtual single limit cycle by overlapping the limit cycles of the oscillators. Using this virtual limit cycle, we can see the details of the phase pattern switching phenomena. Figure 5 shows the overlapped limit cycles in $v_{k1} - v_{k2}$ plane of the phase pattern switching phenomena shown in Fig. 3. At first, before the pulse stimulation, the oscillators are synchronized with the phase pattern shown in the left-hand side of Fig. 4. In this case, Osc 3 and Osc 4 are at the same point because they synchronized at in-phase, and Osc 1 and Osc 2 are in the different points because these are synchronized at ±120° phase shift with Osc 4 (see Fig. 5 (a) and (b)). Once the pulse is added to Osc 1, the motion of Osc 1 veers off the limit cycle (see Fig. 5 (c)–(f)). When Osc 1 reaches to the limit cycle again, it is near Osc 3 and Osc 4 (see Fig. 5 (g)). Therefore, Osc 1 is entrained by Osc 3 and Osc 4 and, finally, the oscillators are synchronized at the phase

Figure 1: (a) The circuit model. (b) Construction of the subcircuits. (c) Schematic expression of the system.

Figure 2: The control signal of the switch $sw(t)$.

Figure 3: The time waveform for case 1.

Figure 4: The phase pattern switching shown in Fig. 3.
pattern shown in the right-hand side of Fig. 4. While the Osc 1 is veering off the limit cycle, the relative positions of Osc 2–Osc 4 are almost invariant because the motions are moving around the limit cycle with the same speed and the perturbation by the pulse is small for them. From these results, the reason why only the phase of the oscillator which is connected to the stimulation unit can be switched is clarified.

Next, we show the simulation results for the Case 2. Note that the stimulation unit is connected to Osc 4. The time waveform is shown in Fig. 3. In this case, the phase pattern has been switched as shown in Fig. 7. In this case, the phase of Osc 1 is taken as the reference signal. The overlapped limit cycles are shown in Fig. 8. In this case, because the oscillators are not identical, the motions can move with the different speeds until the synchronization is achieved. Therefore, because the relative positions of Osc 1–Osc 3 can be changed during the transient state as well as Osc 4, the phases of the multiple oscillators can be switched by the single stimulation.

From these results using the overlapped limit cycles, we can clarify the difference of the phase pattern switching phenomena between Case 1 and Case 2. These results should be useful when we control the phase patterns of those systems.

4. Conclusions

In this paper, we show the transient analyses for the phase pattern switching phenomena in the pulse-driven star-coupled Wien-bridge oscillators by using the overlapped limit cycles. From these results using the overlapped limit cycles, we can clarify the difference of the
phase pattern switching phenomena between the case with and without the parameter deviations. These results should be useful when we control the phase patterns of those systems when we use those systems as the neural networks and the associative memories.

References


Nonlinear Analysis of Varactor-tuned Voltage-Controlled Oscillators

Antonio Buonomo†, Alessandro Lo Schiavo† and Massimiliano Martucci‡

† Facoltà di Ingegneria, Seconda Università di Napoli
Via Roma 29, Aversa (CE) 81031, Italy
‡ STMicroelectronics, Design Center
Via Remo De Feo 1, Arzano (NA), 80022, Italy
Email: antonio.buonomo@unina2.it

Abstract— A nonlinear analysis of the Voltage Controlled Oscillators, including the nonlinear effect of both the active devices and the varactors, is presented. Simple formulas for predicting the steady-state oscillation and its frequency correction due to the effect of nonlinearities are presented, allowing us to analytically predict the tuning curve of VCOs. Their accuracy was validated by circuit simulations.

1. Introduction

Some recent papers devoted to differential Voltage Controlled Oscillators (VCOs) highlighted that the nonlinearity of varactors produces a remarkable deviation of the oscillation frequency from the resonant frequency of LC-tank, and that this nonlinear effect originates some other undesired effects, such as the AM-to-PM conversion, the phase-noise induced by the current tail modulation, as well as the deviation of frequency tuning curve from the expected one [1]-[3]. In the cited papers, the difficulty of predicting the above deviation, owing to the nonlinear nature of problem, is also pointed out.

From both a theoretical and practical point of view, it is thus useful to develop a nonlinear analysis of VCOs allowing us to completely predict the oscillation characteristics [4],[5]. In this paper, applying the perturbation method developed in [6],[7], we show how to perform this analysis for a typical VCO wherein an accumulation-mode MOS varactor is used as a tuning device. A nonlinear model of VCOs accounting for the nonlinear effect of both the varactors and the active devices is presented, as well as simple formulas for calculating all of the parameters of the steady-state oscillation, and its frequency deviation due to nonlinear effects. The formula for the frequency deviation is used for deriving the tuning curve in an analytical form, which is validated through circuit simulations.

2. Circuit model of VCOs

In this Section, we present a perturbation model of differential VCOs allowing us to predict the nonlinear effect on the steady-state oscillation of both the varactors and the active devices. To this end, we consider, as an example, the MOS differential VCO, shown in Fig. 1, and its equivalent circuit around the dc operating point, shown in Fig. 2.

The active part of the VCO is represented in Fig. 1 with a nonlinear two-terminal, whose voltage-current characteristic is written as \( i = i(v) \), where \( i \) and \( v \) are the increment of the drain current and of the differential voltage \( V = V_1 - V_2 \) evaluated with respect to their values at rest, i.e., \( i = I_d - I_{d1} \), \( v = v_1 - v_2 \). Using simple MOS models, an analytical expression of this characteristic has been recently derived in [5], showing that it can be accurately approximated by through a cubic nonlinearity.

The passive part of the VCO is composed of two tank circuits, whose losses are represented in Fig. 2 by the parallel resistances \( R_1 \) and \( R_2 \). The varactors \( D_1 \) and \( D_2 \) represent either MOS capacitors or reverse-biased p-n junctions (diode varactors), as a varactor can be realized in silicon CMOS technology as standard p-n junction, for example forming a p'/n-well structure in an n-well CMOS process.

![Fig. 1. Circuit diagram of a MOS VCO](image-url)
integrated devices, only the differential capacitance is available in practice, which will be denoted by \( C(V_D) \).

\[
q(v_i) = C_0 v_i + \int_0^C c(v_i) dv_i
\]  

where \( C_0 = C(V_{PD} - V_{TUNE}) \) and \( c(v_i) = C(V_D) - C_0 \).

Assuming \( R_1 = R_2 = R \), it can be shown \([7]\) that the circuit in Fig. 1 is described by a system of two mutually-coupled second-order differential equations, which can be written in the perturbation form

\[
\begin{align*}
\left( \omega^2 D_b^2 + 1 \right) v_1 &= -\varepsilon k_1 \omega D_b \left( kv + f(v_1 - v_2) + v_1 \right) \\
&\quad - \varepsilon k_2 \omega^2 D_b^2 g(v_1) \\
\left( \omega^2 D_b^2 + 1 \right) v_2 &= \varepsilon k_1 \omega D_b \left( kv + f(v_1 - v_2) - v_2 \right) \\
&\quad - \varepsilon k_2 \omega^2 D_b^2 g(v_2).
\end{align*}
\]  

where \( k_1 = \sqrt{L/\omega C_0} \), \( k_2 = 1/\omega C_0 \), \( D_b = d/d\theta \), \( \theta = \omega t \), and \( \omega = \omega_0 (1 + \omega_0 \varepsilon + \cdots) \) is the angular frequency of the periodic oscillation with \( \omega_0 = 1/\sqrt{LC_0} \). \( \varepsilon = 1/R \) denotes a small perturbation parameter. The functions \( f(v) \) and \( g(v) \) are the strictly nonlinear components of \( i = i(v) \) and \( q(v) \), respectively, which we assume as small quantities of order \( \varepsilon \). Further, the linear part \( \alpha \) of the nonlinearity, which compensates the circuit losses, is written in the form \( \alpha = k \varepsilon \).

To obtain a complete analytic model of VCO, it is necessary to define the functions \( q(v) \), using existing nonlinear models of varactors.

Here we consider specifically the case of MOS varactors, and in particular accumulation-mode MOS varactors, which are increasingly used as frequency-tuning elements. They are preferred over diode varactors for VCOs because of the greater linearity of the capacitance-voltage characteristic. Unfortunately, an analytic model of the capacitance-voltage characteristic \( C(V_D) \) is not available. As a rule, this characteristic is known only in a graphical form through capacitance measurements at different voltages. However, it can be always expressed, although approximately, through an analytic relationship. As it is shown in Fig. 3, the best approximation to the characteristic seems to be the function \( C(V_D) = C_m + C_1 \tanh(\beta V_D + \xi) \), where

\[
C_m = \frac{C_{\max} + C_{\min}}{2}, \quad C_1 = \frac{C_{\max} - C_{\min}}{2}
\]  

and the remaining parameters are identified as follows. The parameter \( \beta \) is equal to \( mC_1 \), where \( m \) denotes the derivative in the straight line of the curve. The parameter \( \xi \) takes into account the shift of the characteristic curve with respect to the origin \( (V_D = 0) \), and it is usually negative. Both the parameters can be identified graphically.

From (1) it follows that

\[
c(v_i) = -C_a v_i + C_1 \tanh(\beta v_i + \eta)
\]  

\[
e_g(v_i) = -C_a v_i + C_1 \ln[\cosh(\beta v_i + \eta)] - k
\]  

\[
C_a = -C_1 \tanh(\eta), \quad k = C_1 \ln[\cosh(\eta)]/\beta, \quad \eta = \beta V_0 + \xi.
\]  

3. The steady-state oscillation

The steady-state periodic oscillation of the circuit in Fig. 1 is obtained finding the periodic solution of the system (2)-(3). A variety of analytical methods, so-called of the small parameter, can be used to find an approximate analytic solution of it \([8]\). Instead, it does not seems that...
the method of regular perturbation has never applied to similar problems. Using the perturbation method presented in [6], we developed a calculation procedure of the periodic solution of (2), (3), which combines the perturbation method and the Harmonic Balance method incorporating what are probably the main advantages of these classical methods and, at the same time, avoiding what are probably the main shortcomings of each.

According to this calculation procedure, the first-order approximation to the periodic solution is

\[
v_1 = A_0 \cos \vartheta + \tilde{A}_0 h_1(\vartheta) \]

\[
v_2 = \tilde{A}_0 \cos \vartheta + \tilde{h}_1(\vartheta) \]

\[
\vartheta = \omega_0 (1 + \omega_1 \varepsilon) \]  

It can be shown that \( A_0 \) is found solving the nonlinear equation

\[
F_{1,0}(2A_0) = -(2k + 1) \]

\[
\text{while the nonlinear correction to the angular frequency is given in the following explicit form}
\]

\[
\omega_0 = -\frac{k_2}{4A_0} (G_{1,0} - \tilde{G}_{1,0}) \]

where \( F_{1,0} \) and \( G_{1,0} \) denote the Fourier coefficients of the first harmonic of the periodic functions \( f\{ (A_0 - \tilde{A}_0) \cos \vartheta \} \) and \( g\{ A_0 \cos \vartheta \} \), respectively, and \( \tilde{G}_{1,0} = G_{1,0}(-A_0) \), as the condition \( A_0 = -\tilde{A}_0 \) holds [7]. Further, the complex amplitudes of higher-harmonics of periodic oscillation are given by

\[
H_{m,1} = \frac{jmk_1 F_{m,0} - m^2 k_1 G_{m,0}}{m^2 - 1} \]

\[
\tilde{H}_{m,1} = \frac{jmk_1 F_{m,0} + m^2 k_1 \tilde{G}_{m,0}}{m^2 - 1} \]

where

\[
F_{m,0} = \frac{1}{\pi} \int_{0}^{2\pi} f\{ (A_0 - \tilde{A}_0) \cos \vartheta \} \cos m \vartheta \, d\vartheta \]

\[
G_{m,0} = \frac{1}{\pi} \int_{0}^{2\pi} g\{ A_0 \cos \vartheta \} \cos m \vartheta \, d\vartheta \]

From (8)-(11) we can derive formulas for the fundamental harmonic, \( A_0 \), the angular frequency, \( \omega_0 \), and the higher-order harmonics, \( H_{m,1} \) and \( \tilde{H}_{m,1} \), of the steady-state oscillation as a function of circuit parameters, which are useful in both analysis and design of MOS VCOs. As it is easy to recognize, these formulas become very expressive if the approximation \( i = \alpha \omega + c \omega^3 \) is used, which agree well the actual characteristic \( i(v) \) [5].

4. The tuning curve

The usefulness of formulas (8)-(11) is evident. In particular, (9) allows us to obtain the deviation of the oscillation frequency from the ideal one, \( \omega_0 \), in explicit form as a function of the circuit parameters and of the control voltage \( V_{TUNE} \). To show this, we observe that, by virtue of (6), we have

\[
G_{1,0} - \tilde{G}_{1,0} = \frac{1}{\pi} \int_{0}^{2\pi} g\{ A_0 \cos \vartheta \} - g\{ -A_0 \cos \vartheta \} \cos \vartheta \, d\vartheta \]

\[
= - \frac{2C_A A_0}{\pi} \int_{0}^{2\pi} \cos^2 \vartheta \, d\vartheta + \frac{C_1}{\pi \beta} \int_{0}^{2\pi} \cos \vartheta \, d\vartheta \]

where we have put

\[
p(v) = \ln[\cosh(\beta v + \eta)] - \ln[\cosh(-\beta v + \eta)].
\]

To calculate the last integral, we observe that the function \( p(v) \) cannot be integrated with elementary functions. However, an accurate approximation to this function can be obtained using the piece-wise linear function

\[
p(v) = \begin{cases} 
-2 \eta & v \geq v_p \\
(2 \beta \tanh \eta) v - 2 \eta & -v_p \leq v \leq v_p \\
2 \eta & v \leq -v_p 
\end{cases}
\]

where \( v_p = -\eta / \beta \tanh \eta \), as it can be deduced observing that \( p(v) \) reduces to zero in \( v = 0 \), where the derivative is \( 2\beta \tanh \eta \), and it varies between the two asymptotic values \( \pm 2 \eta \).

Using (15), the second integral in the right-hand side of (14) can be solved in a closed-form obtaining

\[
- \frac{4 \omega_0}{\beta} \frac{C_1}{\pi} \left( \sqrt{1 - x^2} - \frac{1}{x} \arcsin(x) \right)
\]

where we set \( x = v_p / A_0 \). As the first integral in the right-hand side of (14) is equal to \( -2C_A A_0 \), from (9) we deduce that
\[
\omega_0 = \frac{C_m}{C_0} \left[ \frac{1}{2} - \frac{x}{\pi} \left( \sqrt{1-x^2} + \frac{1}{x} \arcsin(x) \right) \right]
\]  \tag{17}

and, consequently, that the oscillation frequency in the steady-state is

\[
f = \frac{1}{2\pi} \sqrt{L(C_m + C_1 \tanh(\eta))} \cdot \left[ 1 + \frac{C_m}{C_0} \left( \frac{1}{2} - \frac{x}{\pi} \left( \sqrt{1-x^2} + \frac{1}{x} \arcsin(x) \right) \right) \right].
\]  \tag{18}

Note that, formula (18) yields in an explicit form the relationship between the varactor characteristic and the tuning frequency. Note also that a similar relationship may be derived for any varactor device, on the condition that an analytical model of the \( C(V_D) \) characteristic is available.

Formula (18) has been applied to the calculation of the tuning curve of the circuit in Fig. 1, for which we assumed the following circuit parameters: \( L = 2 \, \text{nH} \), \( R = 400 \, \Omega \), \( V_{DD} = 3 \, \text{V} \), \( I_0 = 3.5 \, \text{mA} \) and the MOS parameters: \( k' = 20 \, \mu A/V^2 \), \( V_T = 0.5 \, \text{V} \), \( W/L = 480 \). The varactors used are pMOS devices realized in a 0.25-\( \mu \)m CMOS technology, operating from the accumulation to depletion mode [2]. The parameter identification yields \( C_m = 0.97 \, \text{pF} \), \( C_1 = 0.3 \, \text{pF} \), \( \beta = 3.5 \, \text{V}^{-1} \) and \( \xi = -0.55 \). The tuning characteristic of the oscillator calculated through formula (18) was reported in Fig. 4 (solid line). It is plotted again \( V_{TUNE} - V_{DD} \), with \( V_{DD} = 3 \, \text{V} \).

The analytical results presented are validated by circuit simulations performed using SPICE. Numerical results agree very well with nonlinear analysis presented in this paper. The tuning characteristic predicted by (18) agrees well with the numerical one (dashed-line), and the slopes of the two curves match notably well in a wide range of the control voltage. The Fig. 4 also reports the ideal tuning curve (dashed-dotted line) calculated using the varactor characteristic, i.e. \( \omega_0 = 1/\sqrt{L(C_{DD} - V_{TUNE})} \), which is much steeper than the numerical and analytical curves.

5. Conclusions

Starting from a suitable circuit model of differential VCOs, which consists of two mutually-coupled nonlinear differential equations, we presented simple formulas for predicting their nearly-sinusoidal oscillation. The model includes the nonlinearities of both the active devices and the varactors. The above formulas, which are derived solving the circuit equations through the solution method presented in [6], [7], are useful for the analysis and the design of VCOs, as well as for studying the nonlinear effect of varactors.

Finally, it should be pointed out that the presented analysis can be easily adapted to cover other LC-tuned topologies.

References

Shooting SOM and its Application for Clustering
Masato Tomita†, Haruna Matsushita† and Yoshifumi Nishio†
Tokushima University
2-1 Minami-Josanjima, Tokushima, Japan
Phone: +81-88-656-7470, FAX:+81-88-656-7471
Email: {tomita, haruna, nishio}@ee.tokushima-u.ac.jp

Abstract—The Self-Organizing Map (SOM) is a popular algorithm for unsupervised learning and visualization introduced by Teuvo Kohonen. One of the most attractive applications of SOM is clustering and several algorithms for various kinds of clustering problems have been reported and investigated. In this study, we propose a new type of SOM algorithm, which is called Shooting SOM (SSOM) algorithm. The important feature of SSOM is that the neurons move like aiming at a target, namely, some neurons find near the center of the area where input data are concentrated and other neurons get away from the former neurons. Because of this feature, SSOM tends to self-organize only in the area where input data are concentrated. We investigate the behavior of SSOM and apply SSOM to clustering problems.

1. Introduction

In data mining, clustering is one of the most important issues and is also very useful for many applications, such as industrial engineering, image processing, biology and medicine. Then, the Self-Organizing Map (SOM) attracts attentions for clustering in recent years. SOM is popular tools for clustering and visualization of high-dimensional data [1]. SOM is an unsupervised neural network introduced by Kohonen in 1982 [2] and is a model simplifying self-organization process of the brain. SOM maps multidimensional data onto a 2-dimensional grid. SOM can classify input data according to similarities and patterns which are obtained by the distance between neurons and is applied to wide fields of data classifications. Although many methods to extract clusters by using SOM have been proposed [3]-[10], it seems to be very difficult to construct a simple method using SOM for universal input data. On the one hand, in the world, the amount and complexity of data increase from year to year and control of large-volume data get a lot of attention. Therefore, it is important to investigate various extraction methods of some clusters from data including some noises.

In this study, we propose a new type of SOM algorithm, which is called Shooting SOM (SSOM) algorithm. The important feature of SSOM is that the neurons move like aiming at a target, namely, some neurons find near the center of the area where input data are concentrated and other neurons get away from the former neurons. Because of this feature, SSOM tends to self-organize only in the area where input data are concentrated. We investigate the behavior of SSOM and apply SSOM to clustering problems.

2. Self-Organizing Map (SOM)

In this section we introduce the conventional SOM algorithm in order to make clear new points of the proposed algorithm. SOM has two-layer structure of the input layer and the competitive layer. In the competitive layer, m neurons are arranged as a regular 2-dimensional grid. The range of the elements of d-dimensional input data \( x_j = (x_{j1}, x_{j2}, \cdots, x_{jd}) \) \((j = 1, 2, \cdots, N)\) are assumed.

(SOM1) The initial values of all the weight vectors \( w_i = (w_{i1}, w_{i2}, \cdots, w_{id}) \) \((i = 1, 2, \cdots, m)\) of the neurons are given between 0 and 1 in a reticular pattern.

(SOM2) An input data \( x_j \) is inputted to all the neurons at the same time in parallel.

(SOM3) We find the winner neuron \( c \) by calculating the distances between \( x_j \) and \( w_i \) according to;

\[
c = \arg \min_i \{ \| w_i - x_j \| \}.
\]

(1)

In other words, the winner neuron \( c \) is the neuron with the weight vector nearest to the input vector \( x_j \). In this study, Euclidean distance is used for Eq. (1).

(SOM4) The weight vectors of all the neurons are updated as;

\[
w_i(t + 1) = w_i(t) + h_{c,i}(t)(x_j - w_i(t)),
\]

(2)

where \( t \) is the learning step. The function \( h_{c,i}(t) \) is called the neighborhood function and it is described.
as follows:

\[ h_{c,i}(t) = \alpha(t) \exp \left( -\frac{\|r_i - r_c\|^2}{2\sigma^2(t)} \right), \]  

(3)

where \( \alpha(t) \) is the learning rate, \( r_i \) and \( r_c \) are the vectorial locations of the neurons on the display grid, and \( \sigma(t) \) corresponds to the widths of the neighborhood function. \( \alpha(t) \) and \( \sigma(t) \) decrease with time according to the following equations:

\[ \alpha(t) = \alpha(0) \left(1 - \frac{t}{T}\right), \]
\[ \sigma(t) = \sigma(0) \left(1 - \frac{t}{T}\right), \]  

(4)

where \( T \) is the maximum number of the learning.

The steps from (SOM2) to (SOM4) are repeated for all the input data, namely, from \( j = 1 \) to \( j = N \).

### 3. Shooting SOM (SSOM)

In this section we introduce SSOM algorithm in order to make clear new points of the proposed algorithm.

#### 3.1. Learning Algorithm

We explain the learning algorithm of SSOM. In SSOM, \( m \) neurons are arranged as a regular 2-dimensional grid. The range of the elements of \( d \)-dimensional input data \( x_j = (x_{j1}, x_{j2}, \ldots, x_{jd}) \) (\( j = 1, 2, \ldots, N \)) are assumed.

(S SOM1) The initial values of all the weight vectors \( w_i = (w_{i1}, w_{i2}, \ldots, w_{id}) \) (\( i = 1, 2, \ldots, m \)) of the neurons are given between 0 and 1 in a reticular pattern.

(SSOM2) An input data \( x_j \) is inputted to all the neurons at the same time in parallel.

(SSOM3) We find the winner neuron \( c \) by calculating the distances between \( x_j \) and \( w_i \) according to:

\[ c = \arg \min_i \{\|w_i - x_j\|\}. \]  

(5)

In other words, the winner neuron \( c \) is the neuron with the weight vector nearest to the input vector \( x_j \). In this study, Euclidean distance is used for Eq. (5).

(SSOM4) We measure whether the winner neuron or 1-neighborhood of the winner neuron or otherwise and show in Fig. 1.

Furthermore, we determine the update rate \( A_i \) for each cases as follows:

\[ A_i = \begin{cases} 
1, & \text{if } i = c \\
-0.04, & \text{if } i \text{ is 1-neighborhood of } c \\
0, & \text{otherwise}.
\end{cases} \]  

(6)

The steps from (SSOM2) to (SSOM5) are repeated for all the input data, namely, from \( j = 1 \) to \( j = N \).

### 4. Simulation Results

#### 4.1. Behavior of SSOM

We consider the 2-dimensional input data of 480 points whose distribution is as Fig. 2. Each 10 clusters distribute 40 points at random. The remaining 80 points are uniformly distributed between 0 and 1 at random. We consider the conventional SOM and the proposed SSOM with 25 neurons.
The parameters for the learning of the conventional SOM and SSOM are chosen as follows;

\[ \alpha(0) = 0.4, \quad \sigma(0) = 1/16. \]  

We execute the learning for all input data once. The simulation results of SSOM is shown in Fig. 3(c). SSOM self-organizes near the center of the area where the input data are concentrated.

Let us examine the behavior of SSOM in more detail. As we can see from Fig. 3(b) and Fig. 3(c), the conventional SOM self-organizes input data including some noises with all neurons. However, we can see that SSOM moves like aiming at a target from learning process. Certain neuron moves toward the cluster to hit a center of the area where input data are concentrated, and other neurons moves toward other cluster and go away. Therefore, only one neuron by the one cluster self-organizes its cluster and SSOM can effectively self-organize many clusters. In doing so, SSOM can self-organize as well as the conventional SOM with minimum neurons because SSOM does not use extra neurons.

4.2. Data Extraction

The concept using SSOM can be exploited to extract the data only in clusters of the input data including some noises, because SSOM can find near the center of such areas by itself. We carry out the extraction of cluster after Fig. 3(d). The extraction method is a relatively simple as follows. After learning, the input data which is within a radius of \( R \) from all neurons on the map are classified into the cluster and extract only the cluster.

The simulation results of SSOM are shown in Figs. 4(a) and (b), respectively \( (R = 0.06) \). As we can see from these figures, SSOM can successfully find near the center of the cluster and can extract as well as the conventional SOM with less neurons.

4.3. Simulation Results for Large Input Data

We consider the 2-dimensional input data of 1240 points whose distribution is as Fig. 5. Each 31 clusters distribute 40 points at random. We consider SSOM with 64 neurons \((8 \times 8)\). The parameters for the learning of SSOM are chosen as follows;

\[ \alpha(0) = 0.3, \quad \sigma(0) = 1/16. \]

We execute the learning for all input data once. The simulation results of SSOM is shown in Fig. 6(a). SSOM self-organizes near the center of the area where the input data are concentrated. Furthermore, we carry out the extraction of cluster after Fig. 6(a). The extraction results of SSOM are shown in Figs. 6(b), \( (R = 0.06) \). As we can see from these figures, SSOM can successfully find the cluster.

5. Conclusions

In this study, we have proposed the Shooting SOM (SSOM). We have explained the differences between the conventional SOM and SSOM have investigated its behavior. Furthermore, we have applied SSOM to extract all near the center of clusters and have confirmed its efficiency.
In the future, we try to discover new applications of SSOM in diverse fields such as sound data processing.

References


Abstract
A novel Logic Oriented Artificial Neural Network (LOANN) is proposed for digital systems. In this paper, many new LOANN of binary flip-flops and registers are designed. The LOANN aim is to design a complete digital neuron computer. The neuro-algebra to design LOANN of any logic function is introduced, where the neuron model forms a complete system with any logic set. The current mode LSI circuits using MOSFETs to implement the LOANNs are also designed and introduced. The novel LOANN has the property that a simple realization to any binary digital system can be achieved. A co-evolutionary algorithm for multi-criterion optimization is used to automate the design of the LOANN. The LOANN simulation results demonstrated the feasibility and the reliability operations with very high performances.

1. Introduction
Artificial neural networks are receiving widespread attentions as potential new architectures for computing systems. Many models have been reported concerning neural networks. However, the research on neural network model has focused mainly on theoretical studies and computer simulations. Several groups have initiated experiments with VLSI implementations and demonstrated a few functional circuits [1-3]. Most of the circuits built so far are relatively small and the designs are also exploratory. The major problem with neural networks is the complicated digital systems designs, especially the sequential logic circuits. Thus, this paper proposes a novel Logic Oriented Artificial Neural Network (LOANN) for digital systems, where we emphasized on the design of many binary flip-flops and registers.

The mathematical properties of LOANN and the model neuron are introduced, where they form a complete system with any logic set. One of the major advantages is the links’ weights and thresholds that are all limited to real number in order to facilitate neural network optimization and hardware implementation. A co-evolutionary algorithm for multi-criterion optimization [4] is used to automate the design of the LOANN (number of hidden layers, interconnections, as well as connections weights). The current mode circuits using MOSFETs [6, 7] to implement the LOANN are also designed and introduced. The simulation results demonstrated the feasibility and the reliability operations with very high performances.

2. Mathematical Properties
Any subset of a set of r truth values \( E(r) \equiv \{0, 1, \ldots, r-1\} \) is called a set-value [8, 9]. In the 2-valued logic system, the finite set of logic states is defined as \( E(2) = \{0, 1\} \).

Figure 1 shows the neuron model. The input transmitted to neuron through these weights may come from other neurons or from external sources. The weighted inputs to a neuron are accumulated and then passed to an activation function which determines the neuron response.

\[ f(Z) = \begin{cases} 1 & \text{if } Z > 0 \\ 0 & \text{if } Z \leq 0 \end{cases} \]

The LOANN structure of 2-valued logic is defined as \( NNP=(G, E(2), f(Z)) \)

Where, \( G=(N, L, W) \),
\( N \): set of nodes (processing elements),
\( L \): set of links (connections),
\( W \): set of synapses’ weights,
\( f(Z) \): Output of the processing element defined as

\[ Z = \sum x_i w_i - \theta \]

Where, \( x_i \): input signals \( x_i \in \{0, 1\} \)
\( w_i \): coefficient weights \( i = 1, 2, \ldots, n \)
\( \theta \): threshold
3. LOANN Applications

Artificial neural networks are receiving widespread attentions as potential new architectures for computing systems [1, 2]. Many models have been reported concerning neural networks. The major problem with neural networks is the complicated digital systems designs, especially the sequential logic circuits [3]. Thus, this section proposes many LOANN novels for digital systems such as SR flip-flops, clocked SR flip-flop, D latch, and buffer registers. A co-evolutionary algorithm [3, 4] for multi-criterion optimization is used to automate the design of the proposed LOANN (number of hidden layers, interconnections, as well as connections weights). Because of space limitation, this section addresses two designs of large-scale integrated circuit for two LOANNs.

SR flip-flop:
The SR flip-flop is a data storage circuit that can be constructed using the basic neuron. The first LOANN of the SR flip-flop is presented in Figure 2(a). The network has two inputs: the set S and the reset R; and two outputs Q and Q'. The characteristic table is presented in Table 1. Figure 2(b) shows the simulation result that demonstrates the functionality, and Figure 2(c) shows the LSI circuit for the LOANN SR flip-flop of Figure 2(a).

Table 1: SR flip-flop characteristic.

<table>
<thead>
<tr>
<th>S</th>
<th>R</th>
<th>Q(T+1)</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>Q(t)</td>
<td>No change</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>Reset</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Set</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>?</td>
<td>Undefined</td>
</tr>
</tbody>
</table>

Clock SR flip-flop:
The LOANN of the clocked SR flip-flop is presented in Figure 3(a). The characteristic table is presented in Table 2. The clock pulse must remain high long enough to allow the state of the flip-flop to change, and the clocked flip flop is level triggered. Figure 3(b) shows the simulation result that demonstrates the functionality.

Table 2: Gated SR flip-flop characteristic.

<table>
<thead>
<tr>
<th>G</th>
<th>S</th>
<th>R</th>
<th>Q(T+1)</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>x</td>
<td>x</td>
<td>Q(t)</td>
<td>No change</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Q(t)</td>
<td>No change</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>Reset</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>Set</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>Undefined</td>
</tr>
</tbody>
</table>

x: Don’t care
D Latch:
The LOANN of the D latch is presented in Figure 4. The D latch (Data flip-flop) receives its designation from its ability to hold data in its internal storage. The binary information present at the data input of the D latch is transferred to the Q output when the control input (CLK) is enabled. The Figure 4(b) shows the simulation result of the D latch network, where the simulation demonstrates the functionality. Figure 4(c) shows the LSI circuit for the LOANN of Figure 4(a).

Table 1: D flip-flop characteristic

<table>
<thead>
<tr>
<th>CLK</th>
<th>D</th>
<th>Q(T+1)</th>
<th>Operation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>x</td>
<td>Q(t)</td>
<td>No change</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>Reset</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>Set</td>
</tr>
</tbody>
</table>

Buffer Register:

Now that we designed the LOANN flip-flops and latches, we can design sequential logic circuits and registers. The simplest register (Figure 5) is the buffer register that consists of four LOANN D latches. The common clock input triggers the system on the high level of each pulse, and the binary data available at the D inputs are transferred into the register. The Q outputs can be sampled to obtain the binary information stored in the register.

Controlled Register with Parallel Load:

This register is controlled with a master clock generator that supplies a continuous train of clock pulses. Also, the register is designed with another control input load signal (L). A novel LOANN of this register is designed as shown in Figure 6(a). The L input to the register determines the action to be taken with each clock pulse. The L signal determines whether the next pulse will accept new information (L=1), or leave the information in the register intact (L=0). Figure 6(b) shows the simulation result, which demonstrates the functionality of the register.

Experimental Simulations:

In all simulations, it is assumed that it takes some time (delay) for a neuron to process its input; and for simplicity this delay is constant for all neurons in a network. As a result of these delays two main issues arise:

1- The output response is delayed from input change. The delay is proportional to the path length between the input and output nodes [5].

2- Glitches (false outputs) appear at the output of some networks. Notice the presence of some glitches in the plot of the output response of respective networks. These transient false outputs are due to the presence of different paths of different lengths (different propagation delays) between input and output. Plots in other figures contain no glitches, since the output nodes of corresponding networks...
are connected to input nodes through paths of the same length. Any network can be made “glitches free” be introducing dummy nodes that have no processing power. This makes the path lengths between the input and output equal and eliminates false output.

4. Conclusion

This paper has proposed a novel LOANN to design many logic systems. The LOANN aim is to design a complete digital neuron computer. The proposed neuron-algebra model forms a complete system with two-valued logic. Thus, the neuro-algebra can design the LOANN of any binary logic function. A co-evolutionary algorithm for multi-criterion optimization is used to automate the design of LOANN (number of hidden layers, interconnections, as well as connections weights). Note that and for the future, the neuro-algebra concept will be extended to design complicated circuits of multiple-valued logic systems. The LOANN of the binary flip-flops and registers are proposed for the first time. The current mode LSI circuits using MOSFETs to implement the LOANN are also designed. The simulation results show that the flip-flops and the registers demonstrated the functionality and the reliability operations. All these LOANN are simulated with MATLAB Simulink.

References:


Reunifying Self-Organizing Map

Haruna Matsushita† and Yoshifumi Nishio†
†Department of Electrical and Electronic Engineering, Tokushima University
Email: {haruna, nishio}@ee.tokushima-u.ac.jp

Abstract—The Self-Organizing Map (SOM) attracts attentions for clustering in these years. In this study, we propose a Reunifying Self-Organizing Map which is a new SOM algorithm. The initial state of all neurons of the proposed SOM are connected to no neuron. However, the neurons are connected gradually to other neurons as learning progressed. The behavior of the reunifying SOM is investigated with application to clustering. We can confirm that the result of using the reunifying SOM includes no inactive neuron.

1. Introduction

Since we can accumulate a huge amount of data in these years, it is important to investigate various clustering methods. Then, the Self-Organizing Map (SOM) attracts attentions in recent years. SOM is an unsupervised neural network introduced by Kohonen in 1982 [1] and is a model simplifying self-organization process of the brain. SOM obtains statistical feature of input data and is applied to a wide field of data classifications [2]-[5]. We can obtain the map reflecting the distribution state of input data using SOM. However, if we apply SOM to clustering of the input data which includes some clusters located at distant locations, there are some inactive neurons between clusters. Because inactive neurons are on a part without the input data, we are misled into thinking that there are some input data between clusters.

In this study, we propose a new SOM algorithm which is the Reunifying Self-Organizing Map. The initial state of all neurons of the proposed SOM are connected to no neuron. However, each neuron has its own physical location on the 2-dimensional grid, so, the neurons are connected gradually to other neurons as learning progressed; hence, the name “reunifying”. In Section II, we explain the learning algorithm of the conventional SOM. In Section III, the learning algorithm of the reunifying SOM in detail. The learning behaviors of the reunifying SOM are investigated in Section IV with applications to clustering of various 2-dimensional input data. Furthermore, clustering ability is evaluated by visualization of results. We can see that there are no inactive neurons using the reunifying SOM.

2. Self-Organizing Map (SOM)

We explain the learning algorithm of the conventional SOM. SOM consists of \( m \) neurons located at a regular low-dimensional grid, usually a 2-D grid. The basic SOM algorithm is iterative. Each neuron \( i \) has a \( d \)-dimensional weight vector \( \mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{id}) \) \((i = 1, 2, \ldots, m)\). The initial values of all the weight vectors are given over the input space at random. The range of the elements of \( d \)-dimensional input data \( \mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jd}) \) \((j = 1, 2, \ldots, N)\) are assumed to be from 0 to 1.

(SOM1) An input vector \( \mathbf{x}_j \) is inputted to all the neurons at the same time in parallel.

(SOM2) Distances between \( \mathbf{x}_j \) and all the weight vectors are calculated. The winner, denoted by \( c \), is the neuron with the weight vector closest to the input vector \( \mathbf{x}_j \);

\[
  c = \arg \min_i \{ \| \mathbf{w}_i - \mathbf{x}_j \| \},
\]

where \( \| \cdot \| \) is the Euclidean distance.

(SOM3) The weight vectors of the neurons are updated as;

\[
  \mathbf{w}_i(t + 1) = \mathbf{w}_i(t) + h_{c,i}(t)(\mathbf{x}_j - \mathbf{w}_i(t)),
\]

where \( t \) is the learning step. \( h_{c,i}(t) \) is called the neighborhood function and is described as a Gaussian function;

\[
  h_{c,i}(t) = \alpha(t) \exp \left( -\frac{\| \mathbf{r}_c - \mathbf{r}_i \|^2}{2\sigma^2(t)} \right),
\]

where \( \| \mathbf{r}_c - \mathbf{r}_i \| \) is the distance between map nodes \( c \) and \( i \) on the map grid, \( \alpha(t) \) is the learning rate, and \( \sigma(t) \) corresponds to the width of the neighborhood function. Both \( \alpha(t) \) and \( \sigma(t) \) decrease monotonically with time as follows;

\[
  \alpha(t) = \alpha(0)(1 - t/T), \quad \sigma(t) = \sigma(0)(1 - t/T),
\]

where \( T \) is the maximum number of the learning.

(SOM4) The steps from (SOM1) to (SOM3) are repeated for all the input data.

3. Reunifying Self-Organizing Map

In this study, we propose a new algorithm of SOM, Reunifying Self-Organizing Map. The initial state of all neurons of the reunifying SOM are connected to no neuron, but each neurons have its own physical location on the 2-D grid. The initial value of the weight vectors are given at orderly position.

(RSOM1) An input data \( \mathbf{x}_j \) is inputted to all the neurons at the same time in parallel.

(RSOM2) The winner \( c \) is found according to Eq. (1).

(RSOM3) The neighborhood distance between \( c \) and each neuron \( i \), denoted by \( n_{c,i} \), is calculated. The
neighborhood distances are defined as shortest-path distances between connected map nodes as Fig. 1. If a neuron \( i \) is not connected directly or indirectly to the winner neuron \( c \), \( n_{c,i} \) is equal to the number of neurons \( m \).

(RSOM4) The weight vectors of the neurons are updated as:

\[
\mathbf{w}_i(t + 1) = \mathbf{w}_i(t) + h_{Rc,i}(t) (\mathbf{x}_j - \mathbf{w}_i(t)),
\]

where \( h_{Rc,i}(t) \) is the neighborhood function of the reunifying SOM and described as:

\[
h_{Rc,i}(t) = \alpha(t) \exp \left( - \frac{(n_{c,i}/m + \| \mathbf{w}_i - \mathbf{x}_j \|)^2}{2\sigma^2(t)} \right).
\]

(RSOM5) A set of 1-neighborhood neurons of winner neuron \( c \), on the assumption that all the neurons are connected (as Fig. 1(a)), is denoted as \( N_{C1} \). The number of \( N_{C1} \) is between two and four when the neurons are located on the 2-D rectangular grid.

A set of the neurons, whose neighborhood distance is the longest in \( N_{C1} \), are denoted as \( S_q \) (as Fig. 2).

\[
S_q = \text{arg max}_i \{ n_{c,i} \}, \quad i \in N_{C1}.
\]

If \( n_{c,q} = 1 \) (as Fig. 2(c)), we perform (RSOM8). However, if not, we perform (RSOM6).

(RSOM6) The connecting neuron \( q \) is chosen from \( S_q \) according to:

\[
q = \text{arg min}_i \{ \| \mathbf{w}_i - \mathbf{x}_j \| \}, \quad i \in S_q.
\]

In other words, the connecting neuron \( q \) is the neuron with the weight vector closest to \( \mathbf{x}_j \) in \( S_q \).

If \( \| \mathbf{w}_q - \mathbf{x}_c \| \leq D(t) \), we perform (RSOM7). However, if not, we except the neuron \( q \) from \( N_{C1} \), and perform the simulation again from (RSOM5).

\( D(t) \) increases monotonically with time according to:

\[
D(t) = \frac{D_{\text{max}}}{T} t + D_{\text{min}}.
\]

(RSOM7) The winner neuron \( c \) is directly connected to a connecting neuron \( q \), namely, \( n_{c,q} \) becomes 1.

(RSOM8) The steps is returned to (RSOM1) and is repeated for all the input data.

4. Application to Clustering

4.1. Simulation 1

First, we consider 2-dimensional input data as shown in Fig. 3(a). The input data is generated artificially as follows. Total number of the input data \( N \) is 600, and the input data include two clusters. 300 data are distributed within a range from 0.2 to 0.8 horizontally and from 0.1 to 0.3 vertically. The remaining 300 data are distributed within a range from 0.2 to 0.8 horizontally and from 0.7 to 0.9 vertically. All the input data are sorted by random.

Both the conventional SOM and the reunifying SOM has \( m = 100 \) neurons \((10 \times 10)\). We repeat the learning 20 times for all input data, namely \( T = 12000 \). The parameters of the learning are chosen as follows:

(For SOM) \( \alpha(0) = 0.5, \sigma(0) = 3 \).

(For Reunifying SOM) \( \alpha(0) = 0.5, \sigma(0) = 0.8, D_{\text{min}} = \sqrt{2}/10, D_{\text{max}} = \sqrt{2}/2 \).

The simulation result of the conventional SOM is shown in Fig. 3(b). We can see that there are some inactive neurons between two clusters.

The other side, the result of the reunifying SOM and its learning process are shown in Fig. 4. We can see from Fig. 4(h) that there are no inactive neurons between two clusters.

4.2. Application to Clustering

(a) Input data. (b) Simulation result of the conventional SOM.
Let us consider the learning process. The initial state of neurons are not connected to other neurons and are located at orderly position as Fig. 4(a). In the early-stage of learning as Figs. 4(b) and (c), we can see that some neurons are connected gradually to other neurons. However, there are some groups of neurons, because the connected neuron is preferentially-selected from the neuron which is connected to no neuron. Furthermore, as Fig. 4(d), there are no neurons which are connected to no neuron, and each two group of neurons self-organize respective two cluster without connecting each other. In the middle stage as Fig. 4(f), two groups of neurons are connected each other according to increase $D(t)$. However, neurons of each group self-organize the respective two clusters without being mingled as Fig. 4(g). in other words, the neurons of each group are not influenced by each other. This is because the learning rate $\alpha(t)$ and the neighborhood radius $\sigma(t)$ decrease according to Eq. (4), namely $\alpha(t)$ and $\sigma(t)$ are small value.

Figure 5 shows distances between neighboring neurons and thus visualizes the cluster structure of the map. Black circles on this figure mean large distance between neighboring map nodes. Clusters are typically uniform areas of white circles. We can see that the boundary line of the reunifying SOM is clear than the conventional SOM.

**4.2. Simulation 2**

Next, we carry out simulation for another 2-dimensional input data example shown in Fig. 6. The input data include two clusters. Total number of input data $N$ is 600. We repeat the learning 20 times for all input data. The parameters of the learning are the same value used in the simulation 1 except for $D_{\text{max}} = \sqrt{2}$.

The results of SOM and the reunifying SOM are shown in Figs. 7(a) and (b), respectively. We can see that there are no inactive neurons in the reunifying SOM although there are some inactive neurons in conventional SOM.
4.3. Simulation 3

Furthermore, we carry out simulation for the concave input data shown in Fig. 8. Total number of input data $N$ is 600. We repeat the learning 20 times for all input data. The parameters of the learning are the same values used in the simulation 1.

Figure 9 shows the result of SOM and the reunifying SOM. As we can see from these figures, the clustering ability of using the reunifying SOM method is effective.

5. Conclusions

In this study, we have proposed the new SOM algorithm which is the Reunifying SOM. The initial state of all neurons of the reunifying SOM are connected to no neuron. However, the neurons are connected gradually to other neurons as learning progressed. We have investigated its behaviors with applications to clustering, and have confirmed the efficiency.
Abstract

The Self-Organizing Map (SOM) is an efficient tool for visualizing high-dimensional data as it performs a topology-preserving projection of the input space on a low-dimensional grid. However, in most of the SOM projection methods, the computation cost is very high, when the size of the data set becomes large. In this paper we present an intuitive and effective SOM projection method with comparatively low computational complexity for the purpose of cluster visualization. This method maps data vectors on the output space based on their responses to different prototype vectors. High-resolution maps can be obtained with a relatively small network size. The proposed method is demonstrated using iris data set.

1. Introduction

In the last decades, many methods have been developed to embed objects, described by high dimensional vectors, into a low-dimensional space. Multi-dimensional Scaling (MDS) is a nonlinear projection method that preserves the inter-point Euclidean distances. The Sammon mapping is a well known MDS technique. The Self-Organizing Maps (SOM), proposed by T. Kohonen [1], is a neural network algorithm based on unsupervised competitive leaning. The SOM creates prototype vectors which have high dimensional value and make them represent same dimensional input data by learning process considering Euclidean distances between input data and prototype vectors. Since each prototype vectors have grid which is in low-dimensional, usually two-dimensional, output space, the SOM can visualize high-dimensional data on the two-dimensional output space. This dimensionality-reducing mapping of the SOM makes inter-relation among the data points and clustering tendency perceptible.

Because of its implicit ability in dimensionality reduction, the SOM has been popularly used as a data clustering and visualization tool. In the SOM the input data is mapped to a low-dimensional regular grid of neurons, which is usually rectangular or hexagonal shaped. The distance between two neurons on the grid indicates the degree of similarity of the data represented by the neurons. Thus, the cluster structure and other patterns of the data can be identified visually from the map created. It is a very intuitive and straightforward way to visualize the data structure. However, in the previous SOM, as the dimension of input data become higher, it becomes more difficult to represent structure of input space accurately. This problem is caused by that the learning neurons which are formed two-dimensional grid have to make itself similar model of input data space even if the dimension of input data is much more than two. Another problem is that the original projection by the SOM alone represents the training results is very crude [2]. Data vectors are mapped to the locations of corresponding best-matching neurons. It is usually difficult to provide much information about the global distribution of the data by observing the resulting row map. For visualizing the data structure, the SOM usually requires assistance from a separate vector projection of the prototype vectors.

In this paper, a new SOM projection method is proposed. First proposition is to use prototype vectors associated with 3D grids in order to represent input data more accurately. Moreover, we use a ranking scheme at the visualization process. The proposed method is demonstrated using iris data set.

2. Visualizing Algorithm

The Self-Organizing Map usually consists of a two-dimensional array of neurons. A prototype vector which has \( n \) attributes (we call this \( n \)-dimensional data) that is \( \mathbf{w}_i = [w_{i1}, w_{i2}, ..., w_{in}] \) is associated with each neuron, where \( n \) is the dimension of the input vectors. At each time step, input vector \( x \) is drawn randomly and presented to the network. This input vector is compared with all the prototype vectors. the nearest prototype vector is called a best matching unit (BMU). A grid number of BMU can be found by calculating the Euclidian distance between the input vector \( x \) and
weight of prototype vector \( w_i \) as follows,

\[
c = \arg\min_{i} \| x - w_i \| \quad (1)
\]

The neighborhood size function is defined to decide the range of learning units. It decreases as time steps. One example of neighborhood size function \( \sigma(t) \) is as follows:

\[
\sigma(t) = d_0 (1 - t/T) , \quad (2)
\]

where \( d_0 \) is a starting width of neighborhood, \( t \) is current time step, and \( T \) is total learning times respectively.

Then the SOM updates the prototype vectors within the neighborhood. The prototype vector \( w_i \) is updated according to

\[
w_i(t + 1) = w_i(t) + h_{ci}(t) [ x(t) - w_i(t) ] , \quad (3)
\]

where \( h_{ci} \) is the time decreasing learning function. A typical smooth neighborhood function could be the Gaussian,

\[
h_{ci}(t) = \alpha(t) e^{-\frac{\| r_c - r_i \|^2}{2\sigma(t)^2}} , \quad (4)
\]

where \( \alpha(t) \) is the learning rate function, \( \| r_c - r_i \| \) is the distance between the winning neuron \( c \), and the neuron \( i \).

The learning process consists of winner selection by equation (1) and adaptation of the prototype vectors by equation (3). After the training has been completed, the map should be topologically ordered so that similar data items are mapped unto nearby map units.

Then, the visualizing process must be carried out in order for the underlying structure of data to be perceived.

3. Proposed visualization scheme

3.1. 3D output map

Recently, we have some problems on the conventional SOM method. One of the most important problems is that prototype vectors are formed on two dimensional grid. In order to project complex structure of data samples on easily understandable output map, the SOM has prototype vectors which are associated with usually one or two dimensioned grid of neurons, and this array represent input space. Since many data sets have more than two attributes, it becomes more and more difficult for two dimensioned array of neurons to represent input data space accurately as the number of attributes of data sample increases. To make output map represent structure of input space more correctly, we adopt 3D array of output map instead of 2D output map.

3.2. Ranking scheme projection

Another problem of the conventional SOM visualization is that the map resolution depends on the size of the map. In the previous SOM, as explained above, input vectors are projected only on the neurons which have nearest value. To achieve a better visualization, the SOM may have been higher number of neurons than that of data vectors. Therefore to get a high-quality result, large number of prototype vectors is required. As a result, learning process becomes impractically time-consuming.

In this paper, we propose a new method which gives a high-resolution map using less prototype vectors. The proposed projection method is based on the standard SOM structure and learning procedure. In visualizing process, we consider not only closest grid, that is BMU, but also the other grids. The response of a data sample to a prototype infers its closeness, which is in turn its membership degree, to a specific unit. Thus a data sample has the highest membership degree to the prototype vector associated with the BMU and it should be mapped to a position closer to the BMU than to other units. There are usually several units with almost as good match as the BMU. Consequently, projecting sample data only on the BMU does not provide accurate information of cluster membership. Intuitively the data item should be projected to somewhere in between the map units with a good match. Analogously, each map unit exerts an attractive force on the data item proportional to its response to that data item. The greater force, the closer the data item attracted to the map unit. The data item will end up in a position where these forces reach equilibrium.

In conventional methods, the SOM projection procedure continues with directly finding the centroid of this spatial response, where the data sample is then mapped. In order to enhance the visual representation, a ranking scheme is used to visualize different degree of cluster membership. First it is required to decide how many units are taken into account. We set
this parameter as $R$.

After that, put order label on each units considering with a distance to sample data which is given by:

- 0 for the closest unit.
- 1 for the second closest unit.
...

$R$ for the $R$th closest unit.

Then, calculate the coordinate $P = (x_1, x_2)^T$ of output map as follows:

$$P = \frac{\sum_{i=0}^{R-1} (\sum_{j=0}^{R-1} d_j - d_i) W_i}{\sum_{i=0}^{R-1} \prod_{j=0}^{R-1} d_j},$$  

(5)

where $d_i$ is Euclidean distance between input vector and weight, $W_i = (y_1, y_2)^T$ is coordinate of the $i$th ranked unit, respectively. Continue to calculate above equation about all the sample data. Then sample data is mapped on coordinate $P$ of the output map. Then SOM processes are summarized as follows:

1. Initializing prototype vectors.
2. Calculate Euclidean distances between prototype vectors and input data.
3. Modify prototype vectors.
4. Set the parameter $R$.
5. Decide the coordinate $P$ of output map by (5).
6. Project the sample data on $P$.

In above flow, 2 and 3 repeat for certain times that is defined by user.

4. Simulation Results

To demonstrate the application of the proposed method, we present the following experiments using Iris Plants data set. The Iris data set is a widely used benchmark in pattern recognition. It contains three classes, where each class refers to a type of Iris plant (Iris-setosa, Iris-versicolor and Iris-virginica). These three classes are divided by their four characteristics (sepal length, sepal width, petal length and petal width). This data set consists of 100 data samples. Before calculation, each element of certain data sample is normalized by the maximum of the category. Thus, all weights of a data sample are between 0 and 1. We use both 2D and 3D-grid prototype vector and ranking scheme to this data set and change rank parameter. Fig. 2 and Fig. 3 show the projection results obtained with the Iris data set using $64(8 \times 8)$ 2D prototype vectors. Fig. 4 and Fig. 5 are 3D projection results using $64(4 \times 4 \times 4)$. Intersection of the grids on the map stand for the prototype vector. Data samples are represented by number according to clusters. "1" stand for Iris-setosa, "2" is associated with Iris-versicolor and Iris-versinica is expressed by "3". We compared maps according to different $R$ values. Notice for $R=1$, where only the BMU is concerned in the projection. Because it can only project input items to map units on a rigid grid, this map does not provide much information about the global shape of the data. However, when $R$ comes to 10, a border of the clusters become more perceptible. Even though four simulations are performed by using same number of prototype vector and weight, the results get better according to the increase of parameter $R$. It also can be said about 3D output map.

Although the obtained 3D maps may be little hard to see the structure of data samples on paper media, on the actual program, the resulting 3D map can be revolved by user so that we can see the relationship between data samples more intuitively.
5. Conclusion

In this paper, a novel approach to visualization techniques for the Self-Organizing Maps are proposed. Our proposed method is simple but effective as shown in result map. Unlike the conventional SOM projection method, which restrict the projection to the junction of the map grid, the proposed method maps the data samples to arbitrary positions across the SOM grid. This enables a high-resolution output map with a comparatively small number of map units. Thus the computational complexity is greatly reduced. And 3D array of prototype vector can make us understand the structure of data samples more intuitively.

The implementation of the proposed method is illustrated using real world high-dimensional data set. The results show the visualization technique has good potential as a tool for structure analysis encountered in high-dimensional input data.

Acknowledgement

This research is supported by the fund of Open Research Center Project from MEXT of Japanese Government (2002-2006)

References

Chaotically Oscillating Sigmoid Function
in Feedforward Neural Network

Yoko Uwate and Yoshifumi Nishio

Dept. of Electrical and Electronics Engineering,
Tokushima University, 2-1 Minami-Josanjima, Tokushima, Japan,
Phone: +81-88-656-7470, Fax: +81-88-656-7471,
Email: {uwate, nishio}@ee.tokushima-u.ac.jp.

Abstract—In this study, we propose the feedforward neural network with chaotically oscillating sigmoid function. By computer simulations, we confirm that the proposed neural network can find good solutions in early time of the back propagation learning process.

1. Introduction

Recently, studies on the human brain have been carried out actively on various levels. Many modelings of the human brain with the visual or the audio sensation are reported [1]-[3] due to development of the brain researches. However, the investigation of modeling of higher functions in the human brain is just getting started. We consider that it is very important to apply these high functional mechanisms of the human brain to novel artificial neural networks.

Back Propagation (BP) learning [4] is one of engineering applications of artificial neural networks. The BP learning operates with a feedforward neural network which is composed of an input layer, a hidden layer and an output layer, and the effectiveness of the BP learning has been confirmed in pattern recognition, system control, signal processing, and so on [5]-[7]. The BP learning process requires that the input-output functions are bounded and differentiable functions. One of the most commonly used functions satisfying these requirements is the sigmoid function. This function is an S shaped monotonic increasing function that has the general form as following equation:

\[ f(x) = \frac{1}{1 + e^{-\varepsilon x}} \] (1)

where \(\varepsilon\) is a constant that determines the steepness of the S shaped curve. Some curves of the function for different values of \(\varepsilon\) are illustrated in Fig. 1.

![Sigmoid function](image1)

Figure 1: Sigmoid function.

The nonlinearity of the sigmoid function has an effect on modifying connection weights and it is very important for BP learning.

We consider that neurons in the human brain do not always output the same output for the same input. In order to reflect this idea to the feedforward neural networks, in this study, we propose the feedforward neural network with chaotically oscillating gradient of the sigmoid function. In order to confirm the effectiveness of the chaotically oscillating sigmoid function, we carry out computer simulations using other shaking methods. Further, we compare the proposed network to the Simulated Annealing (SA) method. By computer simulations, we confirm that the proposed network with chaotically oscillating sigmoid function can find good solutions in early time of the BP learning process.

2. Chaotically Oscillating Sigmoid Function

The BP learning process requires that the input-output functions are bounded and differentiable functions. One of the most commonly used functions satisfying these requirements is the sigmoid function. This function is an S shaped monotonic increasing function that has the general form as following equation:

\[ f(x) = \frac{1}{1 + e^{-\varepsilon x}} \] (1)

where \(\varepsilon\) is a constant that determines the steepness of the S shaped curve. Some curves of the function for different values of \(\varepsilon\) are illustrated in Fig. 1.

![Sigmoid function](image1)

Figure 1: Sigmoid function.

The nonlinearity of the sigmoid function has an effect on modifying connection weights and it is very important for BP learning. We propose the feedforward neural network with chaotically oscillating gradient (\(\varepsilon\)) of the sigmoid function for BP learning.

The authors have investigated the performance of the Hopfield neural network solving combinatorial optimization problems when chaos is inputted to the neurons as noise [8]-[10]. By computer simulations, chaotic noise has been confirmed to gain better performance to escape out of local minima than random noise. Hence, we consider that various features of chaos are effective for neural networks.
The logistic map is used to shake $\varepsilon$ of the sigmoid function chaotically:

$$\dot{\varepsilon}(t + 1) = \alpha \varepsilon(t)(1 - \varepsilon(t)). \quad (2)$$

Varying the parameter $\alpha$, Eq. (2) behaves chaotically via a period-doubling cascade. Further, it is well known that the map produces intermittent bursts just before periodic-windows appear. We apply the sequence generated by the logistic map to the sigmoid function after the following linear transform to set the standard as 1.0 and control the amplitude.

$$\varepsilon(t) = 2A(\varepsilon(t) - 1) + 1 \quad (3)$$

where $A$ corresponds to the range of $\varepsilon$. One example of $A = 0.5$ is shown in Fig. 2.

![Oscillating $\varepsilon$ by logistic map.](image)

Figure 2: Oscillating $\varepsilon$ by logistic map.

### 3. BP Learning Algorithm

The standard BP learning algorithm was introduced in [4]. The BP is the most common learning algorithm for feedforward neural networks. In this study, we use the batch BP learning algorithm. The batch BP learning algorithm is expressed by a formula similar to the standard BP learning algorithm. The difference lies in the timing of the weight. The update of the standard BP is performed after each single input data, while for the batch BP the update is performed after all input data has been processed. The total error $E$ of the network is defined as

$$E = \sum_{p=1}^{P} E_p = \sum_{p=1}^{P} \left\{ \frac{1}{2} \sum_{i=1}^{N} (t_{pi} - o_{pi})^2 \right\}, \quad (4)$$

where $P$ is the number of the input data, $N$ is the number of the neurons in the output layer, $t_{pi}$ denotes the value of the desired target data for the $p$th input data, and $o_{pi}$ denotes the value of the output data for the $p$th input data. The goal of the learning is to set weights between all layers of the network so as to minimize the total error $E$. In order to minimize $E$, the weights are adjusted according to the following equation:

$$w^{k-1,k}_{i,j}(m + 1) = w^{k-1,k}_{i,j}(m) + \sum_{p=1}^{P} \Delta p w^{k-1,k}_{i,j}(m),$$

$$\Delta p w^{k-1,k}_{i,j}(m) = -\eta \frac{\partial E_p}{\partial w^{k-1,k}_{i,j}}, \quad (5)$$

where $w^{k-1,k}_{i,j}$ is the weight between the $i$th neuron of the layer $k - 1$ and the $j$th neuron of the layer $k$, $m$ is the learning time, and $\eta$ is a proportionality factor known as the learning rate. In this study, we add to the second line of Eq.(3) an inertia term, leading to

$$\Delta p w^{k-1,k}_{i,j}(m) = -\eta \frac{\partial E_p}{\partial w^{k-1,k}_{i,j}} + \zeta \Delta p w^{k-1,k}_{i,j}(m-1), \quad (6)$$

where $\zeta$ denotes the inertia rate.

### 4. Simulated Results

We consider the feedforward neural network producing outputs $x^2$ for inputs data $x$ as one learning example. The sampling range of the input data is $[-1.0, 1.0]$ and the step size of the input data is set to be 0.01. We carried out the BP learning by using the following parameters. The learning rate and the inertia rate are fixed as $\eta = 0.2$ and $\zeta = 0.02$, respectively. The initial values of the weights are given between $-1.0$ and $1.0$ at random. The learning time is set to 10000, and the 8 neurons are prepared in the hidden layer. The network structure using this study and learning example are shown in Fig. 3.

![Network structure and learning example.](image)

Figure 3: Network structure and learning example.

#### 4.1. Performance of Learning Process

We investigate the learning efficiency as the total error between the output and the desired target. We define “Average Error $E_{ave}$” by the following equation as mean square error.

$$E_{ave} = \frac{1}{P} \sum_{p=1}^{P} \left\{ \frac{1}{2} (t_p - o_p)^2 \right\} \quad (7)$$

The bifurcation parameter of the logistic map is set to $\alpha = 3.8274$ generating intermittency chaos. The
gradient of the sigmoid function of the conventional network is fixed as $\varepsilon = 1.0$. Figure 4 shows one example of the simulation results when the amplitude of $\varepsilon$ of the sigmoid functions are changed. The horizontal axis is iteration time and the vertical axis is $E_{ave}$. This figure shows three learning curve of the proposed network and the conventional network, respectively, when the initial conditions of the connection weights are changed. From this figure, we can confirm that the proposed network with the chaotically oscillating sigmoid function gains better performance than the conventional network when the amplitude of $\varepsilon$ is set to 0.5 (Fig. 4(a)). The learning curve of the proposed network converges oscillatory. On the other hand, the proposed network shows similar or weak performance to the conventional network when the amplitude of $\varepsilon$ are set to 0.2 and 0.8 (Fig. 4(b) and (c)). We consider that the proposed network gains good performance by using appropriate amplitude of $\varepsilon$ of the chaotically oscillating sigmoid function.

4.2. Comparison of Shaking Methods

In this section, in order to confirm the effectiveness of the chaotically oscillating sigmoid function, we carry out the computer simulations by other shaking methods; fully developed chaos and at random. The fully developed chaos was realized by setting parameter of the logistic map (Eq. (2)) as $\alpha = 4.0000$. The simulation result is shown in Fig. 5. In this figure, the results of the four cases “intermittency chaos,” “fully developed chaos”, “random” and “conventional network” are shown.

We can confirm that the learning curve of the random method is more oscillatory than the intermittency chaos and the fully developed chaos. We consider that the chaotically oscillating $\varepsilon$ is important to find good solution for BP learning.

4.3. Efficient Learning

The proposed network with chaotically oscillating sigmoid function finds better solutions than the conventional network. However, the learning curve of the proposed network is oscillatory and do not converge. We consider that the proposed network can converge effectively, if oscillating $\varepsilon$ stops on BP learning.

Figure 6 shows the simulated result when the stopped iteration time is set to 3000, 4000 and 5000. After oscillating stop, the gradient is fixed as $\varepsilon = 1.0$. From this figure, we can see that the proposed network converges to good solution by oscillating stop. We consider that it is important to oscillate the gradient of the sigmoid function in early time of learning process to find good solution.
4.4. SA Method

Simulated annealing (SA) is a generic probabilistic meta-algorithm for global optimization problems, namely locating a good approximation to the global optimum of a given function in a large search space. SA can find a good solution by decreasing the gradient of the sigmoid function gradually. In this section, we investigate the performance of the SA method and the proposed network with the concept of SA. For comparison, the learning ability of the proposed network and the conventional network are investigated. The changing $\varepsilon$ in these networks are shown in Fig. 7. The horizontal axis is time and the vertical axis is $\varepsilon$. The simulated result is shown in Fig. 8. From this figure, the SA method and the proposed network with SA do not escape from local minima as well as the conventional network. We consider that the network needs some irregular change of neurons themselves in early learning time to find a good solution.

5. Conclusions

In this study, we proposed the feedforward neural network with chaotically oscillating gradient of the sigmoid function. By computer simulations, we confirmed that the feedforward neural network with chaotically oscillating sigmoid function can find good solutions in early time of the BP learning process.

References

Comparative Dynamics for Climate Policy Choice

Yasuo NONAKA

College of Economics, Kanto Gakuin University
1-50-1 Mutsuura-higashi, Kanazawa-Ku, Yokohama 236-8501, Japan
Email: nonaka@kanto-gakuin.ac.jp

Abstract—This paper investigates long-run effects of climate policy. A policymaker introduces global carbon taxes in order to reduce atmospheric CO2 (carbon dioxide) accumulation. This paper considers the interaction between economic activities under taxation and carbon uptake by ecosystem. It numerically demonstrates that given the nonlinear carbon sink, the policymaker faces difficulties for decision-making of the tax rate.

1. Introduction

For rapid global warming, prompt and concrete policy actions are requested. Global warming is caused by increment of atmospheric concentration of carbon dioxide (CO2). The Kyoto protocol that has come into effect in 2005 intends to stabilize the atmospheric CO2 concentration by reducing CO2 emissions from Annex I countries.\footnote{In fact, [2] provides 6 scenarios those are aimed to stabilize atmospheric CO2 concentrations between 450 ppm and 1000 ppm.}

The existing studies analyze the problem of global warming in dynamic game frameworks; see [1], [3], [4], and [6] to name a few. In these studies policy authorities decide the global carbon tax rate for given observation of the atmospheric CO2 accumulation. Then, equilibrium strategies using the concept of differential game framework have been analyzed. The equilibrium strategies are aimed at determining the steady-state levels of CO2 accumulation.

These studies assume a linear state equation of the atmospheric CO2 concentrations. However, since the problem of global warming can be characterized by dynamical system describing a interaction between ecosystem and our economic system, the state equation could be nonlinear. Further, introducing climate policies such as carbon taxes and other policy instruments may cause complex behavior of the atmospheric CO2 concentrations.\footnote{The Kyoto Protocol has set new policy instruments, often referred to as Kyoto mechanism that includes the clean development mechanism (CDM), joint implementation (JI) and international emission trade (IET).}

This paper analyzes interactions between ecosystem and economic system under the global carbon taxation. Then the policy authorities set their political goal to stabilize the system and lead the CO2 accumulation to a level which does not cause rapid climate changes. With nonlinearity of the system the authorities face sharpness of taxation parameter setting.

2. Model

In this section we construct a dynamic model for carbon taxation using a linear state equation for atmospheric CO2 accumulation. Let \( x_t \) denote an observation of CO2 accumulation in period \( t \). A part of CO2 in atmosphere is absorbed by the ecosystem such as terrestrial plants. On the other hand, our economic activities using of fossil fuel will increase the accumulation. Let \( q_t \) denotes the amount of economic activities in period \( t \). Then, the CO2 accumulation equation is given by

\[
   x_{t+1} = (1-\delta)x_t + yq_t, \quad 0 < \gamma, \delta < 1.
\]

Here \( \delta \) is an absorption factor of CO2 by the ecosystem and \( y \) is a conversion rate from CO2 emission by economic activities to atmospheric accumulation.

Policy authorities introduce the global carbon tax to induces the atmospheric CO2 accumulation to a political goal. We consider that the rate of the carbon tax depends on the level of the CO2 accumulation. Then define the carbon tax rate \( \tau(x, \cdot) \) as

\[
   \tau(x, \cdot) \equiv \max[s(x-x_{\text{Ref}}), 0], \quad s > 0.
\]

Here \( x_{\text{Ref}} \) is the reference point of CO2 accumulation at which the policy authorities decide introducing carbon taxes and \( s \) is an exogenous taxation parameter.

In this model production behavior of the representative firm express our economic activities. The profit of the representative firm \( \Pi \) is defined by

\[
   \Pi(x, q) = \alpha q(\beta - q) - \tau(q, x).
\]

The right hand side of (2.3) implies if current CO2 accumulation belows the reference point \( (x \leq x_{\text{Ref}}) \), the firm has a unique profit-maximizing output \( \hat{q} = \beta/2 \). On the other hand, if CO2 accumulation exceeds the reference point \( (x > x_{\text{Ref}}) \), the output of the firm belows \( \hat{q} \). This immediately means a reduction of CO2 emission to the atmosphere.

Solve the profit-maximizing problem of the firm with respect to \( q \). Then the optimal output is a map of CO2 accumulation \( x \) given by

\[
   \phi(x) = \arg \max_q \Pi(x, q)
\]

\[
   = \begin{cases} 
   \hat{q} & (x \leq x_{\text{Ref}} \text{ or } s = 0) \\
   \hat{q} - \frac{s(x - x_{\text{Ref}})}{2\gamma} & (x_{\text{Ref}} < x \leq \hat{x} \text{ and } s > 0) \\
   0 & (x > \hat{x} \text{ and } s > 0).
   \end{cases}
\]
Here \( \hat{x} = x_{\text{ref}} + \alpha \hat{y} / \beta \) is shutdown point for the firm with respect to \( \text{CO}_2 \) accumulation.

The truncated map \( \phi \) given by (2.4) is a reaction function of the representative firm for the observation of atmospheric \( \text{CO}_2 \) accumulation \( x \). Therefore the firm decide its production plan in period \( t + 1 \) for the observation in period \( t \). Then

\[
q_{t+1} = \phi(x_t).
\]

(2.5)

Hereafter we analyse a linear discrete system consists of \( \text{CO}_2 \) accumulation equation (2.1) and the reaction function of the firm (2.5).

3. Stationary state of \( \text{CO}_2 \) accumulation

This section analyzes stationary state of \( \text{CO}_2 \) accumulation and its stabilities. The production output of the representative firm in \( t + 1 \) given by (2.5) is equal to zero if the global carbon tax is introduced (\( s > 0 \)) and then an observation of \( \text{CO}_2 \) accumulation exceeds shutdown point \( \hat{x} \). In this case, the \( \text{CO}_2 \) accumulation will decreases the absorption factor by ecosystem \( \delta (0 < \delta < 1) \). Finally, \( x_t \) belows \( \hat{x} \) and then the firm will restarts the production. Here we assume that \( x_t < \hat{x} \) for arbitrary period \( t \).

If policy authorities did not introduce the carbon tax (\( s = 0 \)), the representative firm choose profit maximizing output \( \hat{q} = \beta / \delta \) in each period. Thus we can rewrite (2.1) as a first order linear difference equation

\[
x_{t+1} = (1 - \delta) x_t + \frac{\gamma \beta}{2}. \tag{3.1}
\]

(3.1) can be easily solved and for initial condition \( x_0 \) its general solution is given by

\[
x_t = \frac{(1 - (1 - \delta) \gamma t)}{2 \delta} + x_0 \tag{3.2}
\]

This implies that \( x_t \) will converge to a unique stationary state \( x^* = \beta \gamma / 2 \delta \) for \( t \to \infty \).

Let \( x_{\max} = \max(x_t) > x_{\text{ref}} \) denote the maximal permissible value of \( \text{CO}_2 \) accumulation which does not cause serious problem for global environment such as rapid climate changes. Suppose that the stationary state without carbon taxation belows the maximal permissible value (i.e. \( x^* \leq x_{\text{ref}} \)). Then we have no reason to introduce the carbon tax. However, consider the reason that we should start immediate and concrete political action for global warming, then it is appropriate to assume that \( x^* < x_{\text{ref}} \).

If policy authorities introduced the global carbon tax (\( s > 0 \)), we have a linear difference system of \( \text{CO}_2 \) accumulation and the production output of representative firm:

\[
x_{t+1} = (1 - \delta) x_t + \frac{\gamma q_t}{s} \tag{3.3}
\]

\[
q_{t+1} = \frac{(x_{\text{ref}} - x_t) + \beta}{2 \alpha} \tag{3.4}
\]

Stationary point of the system (3.3), \((x^*, q^*)\) is always unique if it exists and is given by

\[
(x^*, q^*) = \left( \frac{(x_{\text{ref}} + \alpha \beta) \gamma}{2 \alpha \delta + s \gamma}, \frac{(x_{\text{ref}} + \alpha \beta) \delta}{2 \alpha \delta + s \gamma} \right) \tag{3.4}
\]

Here it is obvious that \( x^*, q^* > 0 \) and stationary point (3.4) always exists. Further, since \( x_{\text{ref}} < \beta \gamma / 2 \delta \), we have

\[
\frac{\partial x^*}{\partial s} = \frac{2 \alpha \delta \gamma - \beta \gamma \gamma}{(2 \alpha \delta + s \gamma)^2} > 0. \tag{3.5}
\]

Thus, for increment of taxation parameter \( s \), the stationary state \( x^* \) will decrease and converges to the reference point \( x_{\text{ref}} \).

2-dimensional linear difference system (3.3) can be rewritten as a second-order linear difference equation:

\[
x_{t+2} = (1 - \delta) x_{t+1} - \frac{\gamma t}{2 \alpha} x_t + \frac{(x_{\text{ref}} + \alpha \beta) \gamma}{2 \alpha} \tag{3.6}
\]

For arbitrary constant \( C_1 \) and \( C_2 \), general solution for (3.6) is given by

\[
x_t = x^* + \lambda_1 C_1 + \lambda_2 C_2, \quad t \geq 0. \tag{3.7}
\]

Here \( \lambda_i, i = 1, 2 \) are eigenvale of Jacobian matrix:

\[
J = \begin{pmatrix}
1 - \delta & \gamma \\
-\frac{\delta}{2 \alpha} & 0
\end{pmatrix} \tag{3.8}
\]

and given by

\[
\lambda_1 = \frac{1 - \delta - \sqrt{\Delta}}{2}, \quad \lambda_2 = \frac{1 - \delta + \sqrt{\Delta}}{2} \tag{3.9}
\]

Here \( \Delta = 4 \alpha^2 (1 - \delta^2) - 8 \gamma \beta / \delta \) is a determinant of eigenvalue \( \lambda_i \).

(3.7) is stable if and only if

\[
\max \{|\lambda_1|, |\lambda_2|\} < 1 \tag{3.10}
\]

Consider the case that \( \Delta \geq 0 \), that is, Jacobian matrix \( J \) has real eigenvalues. If \( s = 0 \), then \( \lambda_1 = 0 \) and \( \lambda_2 = 1 - \delta \). Thus, \( x_t \) converges to the stationary state \( x^* = \gamma \beta / 2 \delta \) for \( t \to \infty \). If \( 0 < s \leq \alpha (1 - \delta^2) / 2 \beta \), then \( 0 < \lambda_1 < 1 - \delta / 2 \) and \( 1 - \delta / 2 < \lambda_2 < 1 \). Again, \( x_t \) converges to \( x^* \) for \( t \to \infty \). Therefore, \( \text{CO}_2 \) accumulation always converges to the stationary state if the system has real eigenvalues.3

On the other hand, if \( s > \alpha (1 - \delta^2) / 2 \beta, \Delta < 0 \) then, \( \lambda_i \) are conjugate complex eigenvalues. In this case, \( x^* \) is asymptotically stable the absolute value of the eigenvalues is less than unity. However the absolute value is greater than unity if \( s > 2 \alpha / \gamma \). Then the stationary state \( x^* \) becomes unstable. Here, \( x_t \) shows periodic behavior around \( x^* \).4

3If \( s > \beta \), then the representative firm will stop its production before \( \text{CO}_2 \) accumulation converges to the stationary state because of high tax rate. However, since

\[
\frac{x - x_t}{s} = \frac{2 \alpha (x_{\text{ref}} + \alpha \beta) \gamma}{s(2 \alpha \delta + s \gamma)}
\]

and it is always positive for finite \( s > 0 \). Therefore, \( x_t \) always converges to \( x^* \).

4For increment of carbon tax rate, the representative firm finally stops its production. However, if \( s_t \) decreases to the reference point \( x_{\text{ref}} \), then the firm starts its production again. However, if \( \text{CO}_2 \) accumulation increases again, then the tax rate increases too and this leads to decrease of the firm’s production output. From the above-mentioned, with high tax rate, the atmospheric \( \text{CO}_2 \) accumulation fluctuates between the reference point \( x_{\text{ref}} \) and the maximum value \( x_{\text{ref}} + \alpha \beta / \delta \) periodically.
4. Comparative Dynamics for carbon taxation

In this section, through numerical simulation, we examine how the asymptotic behavior of the atmospheric CO₂ accumulation varies for introducing the global carbon tax and changes of the tax rate.

First, we set the absorption factor $\delta = 0.005$ and the conversion rate $\gamma = 0.5$. Second, the parameter values of the profit function of the representative firm $\Pi$ are given as $\alpha = 4$ and $\beta = 2$. Here the profit maximizing output $\hat{q} = 1$ without taxation ($s = 0$). Then we can normalize the output of the firm in $[0, 1]$.

Figure 1: Carbon tax rate and CO₂ accumulation.

Figure 1 illustrates the variation of stationary state value of CO₂ accumulation $x^*$ and break even point $\hat{x}$ where the reference value $x_{\text{ref}} = 60$. We can see in this figure that for arbitrary tax rate $s$, $x^* < \hat{x}$ and for increasing $s$, $x^*$ and $\hat{x}$ converge to $x_{\text{ref}}$.

Given other parameters the stationary value $x^*$ loses its stability for $s > 2\alpha / \gamma = 16$. Figure 2 illustrates periodic behavior of CO₂ accumulation around unstable $x^*$. Here we can observe two cases. In Fig. 2(a), CO₂ accumulation has an absorbing barrier at $q = 1$. On the other hand, in Fig. 2(b), CO₂ accumulation has an absorbing barrier at $q = 0$.

Now we consider the optimal carbon tax rate. Under optimal tax rate, the firm maximizes its profit where CO₂ accumulation stays in the political objective range. This implies that under taxation, the stationary value $x^*$ is eventually equal to the maximal permissible value $x_{\text{max}}$. Solving $x^* = x_{\text{max}}$ for $s$ we have the optimal tax rate $s^*$ as

$$
    s^* = \frac{\alpha \beta \gamma - 2x_{\text{max}} \alpha \delta}{x_{\text{max}} - x_{\text{ref}} \gamma}.
$$

Figure 2: Periodic behavior of CO₂ accumulation. ($x_{\text{ref}} < \frac{\beta \gamma}{2x_{\text{max}}}, s > 2\alpha / \gamma$)

5. Nonlinear transitional equation

In above discussion, we have considered the CO₂ absorption factor of by ecosystem is constant. In this reason, the CO₂ accumulation equation is linear and policy authorities can lead the accumulation to political goal by setting the taxation parameter appropriately. However the CO₂ absorption factor depends on various elements in ecosystem, especially CO₂ accumulation level. In this section, we consider that the absorption factor is a map of CO₂ accumulation level.

Suppose that the CO₂ absorption factor $\delta$ will monotonically decreasing when the accumulation level deviate from the reference point. Then $\delta$ is determined by a map $\psi(x)$

$$
    \delta = \psi(x) \equiv \tilde{\delta} \exp(\mu(x - x_{\text{ref}})).
$$

Here $\tilde{\delta}$ is the level of the absorption factor when the accumulation level is equal to the reference point $x_{\text{ref}}$.\(^6\)

\(^{6}\)In the other words, (5.1) means that $\delta$ is exponentially distributed with respect to CO₂ accumulation level.
Substituting (5.1) into (2.1), the dynamical system (3.3) becomes
\[
\begin{align*}
x_{t+1} &= (1 - \delta \exp(\mu (x - x_{\text{ref}}))) x_t + \gamma q_t \\
q_{t+1} &= \frac{s(x_{\text{ref}} - x_t) + \beta}{2}\end{align*}
\]  
(5.2)

(5.2) is a nonlinear system and does not have algebraic solution. Therefore we numerically examine its behaviors. In numerical simulations, we apply same parameter values in Section 4. Further, we consider the linear system (3.3) where \( \delta = \delta^* \) as a comparison.

Through the numerical simulation, we confirmed the stabilities of both systems. However, in the nonlinear system (5.2) CO₂ accumulation level converges to different value depending on its initial conditions \( x_0 \) while it always converges to a unique stationary state \( x^* \) in the linear system (3.3). In the nonlinear system, \( x_{\infty} \) has a threshold value \( x_0 \) with respect to the initial conditions. If \( x_0 < x_0 \), CO₂ accumulation converges to a level less than \( x_{\text{ref}} \). Then the representative firm gets the maximal profit. On the other hands, if \( x_0 > x_0 \), the accumulation converges to the level greater than the stationary state \( x^* \) in the linear system.

![Diagram](image)

Figure 3: Threshold value for the initial conditions. \((\delta = 0.005, \delta = 0.5, \mu = 0.5, s = 0.1)\)

Figure 3 shows the relation between the initial condition \( x_0 \) and convergence \( x_{\infty} \) of CO₂ accumulation in the nonlinear system 5.2 where \( x_{\text{ref}} = 60 \). In Fig. 3, \( x_{\infty} \) converges to the level less than \( x_{\text{ref}} \) if \( x_0 < x_0 \). However, if \( x_0 > x_0 \), \( x_{\infty} \) discontinuously increases. It is determined between stationary state in the linear system \( x^* \) and shutdown point \( \xi \). Eventually CO₂ accumulation level stays the initial condition during \( x_0 \) approaches to \( \xi \).

Further numerical simulation shows that the threshold value of the initial conditions \( x_0 \) and the convergence \( x_{\infty} \) has sharpness with respect to parameter \( \mu \) and \( s \). It implies that under nonlinear system, it is difficult to forecast the convergence level of CO₂ accumulation before setting the taxation parameter \( s \).

6. Concluding remarks

In this paper, we demonstrate that carbon taxation can control the atmospheric CO₂ accumulation if its state equation is linear. However if we introducing nonlinearity to the equation in which the CO₂ absorption factor depends on the accumulation, then the asymptotic behavior of the equation has sharpness with respect to the taxation parameter. This implies that the policy authorities may encounter difficulties to work on climate policy under the nonlinear ecosystem.

References


A Simple Structure of Bifurcating Economies

Junichi Minagawa

†Graduate School of Economics, Chuo University
742-1 Higashi-nakano, Hachioji, Tokyo 192-0393 Japan

Abstract—This paper extends Scarf [11]’s three-commodity economy to an n-commodity economy and presents a structural approach to Scarf’s, Gale [5]’s, and others’ instability results. It is extracted from the extended Scarf example a structure of systems with a bifurcation characterized by the stability change of its equilibrium. The extended bifurcating economy is classified into the Scarf type economy and the Gale-Bala type economy according as the number of commodities.

1. Introduction

In the context of the stability of competitive equilibrium, Scarf [11] showed, implicitly, a bifurcating exchange economy in which the continuous time tâtonnement changes the qualitative type as its parameter passes through a critical value. From the viewpoint of nonlinear dynamics, Bala [2] and Mukherji [8] presented bifurcating exchange economies explicitly, more or less, following Scarf’s result. Bala extended Gale [5]’s example to a parametric class of exchange economy and showed that the tâtonnement exhibits a pitchfork bifurcation. Whereas Mukherji extended Scarf’s first example to a parametric class of exchange economy and showed that the tâtonnement undergoes a Hopf bifurcation.

This paper extends Scarf’s three-commodity economy in the second example to an n-commodity economy in which the excess demand function includes those in Scarf’s, Gale’s, and Bala’s examples, and attempts to structurally characterize these instability results. That is, we point out a structure of systems, extracted from the extended Scarf example, with a bifurcation characterized, regardless of the dimension n, by the stability change of its equilibrium from locally stable to locally completely unstable (all eigenvalues have positive real parts). It is also shown that the analysis of the extended Scarf example gives a particular example for Dierker [3, 4]’s remarks on the number of equilibria of an economy. Finally, the extended exchange economy is classified into the Scarf type bifurcating economy and the Gale-Bala type bifurcating economy according as the number of commodities.

2. Scarf’s Example: The Case of n

Scarf [11] presented examples of instability in a three-commodity exchange economy with a unique competitive equilibrium under the Walrasian tâtonnement (see also Negishi [9] and Anderson et al. [1] for meanings of the examples). Instead of a stable equilibrium, the tâtonnement has a family of closed orbits around the equilibrium in the first example and a closed orbit attracting neighbouring orbits in the second example. While Gale [5] provided a two-commodity exchange economy which has three equilibria and the symmetrical ‘fair’ equilibrium is unstable. In the following, we will extend Scarf’s second example to an n-commodity exchange economy including the original example as a particular case.

Consider an exchange economy in which there are n agents and n commodities and n ≥ 2. Let H = {1, 2, . . . , n} be an index set. Let xj = (xj1, xj2, . . . , xjn) be a commodity vector in which xji is the amount of the jth commodity consumed by the ith agent. Let wj = (ωj1, ωj2, . . . , ωjn) be an initial endowment vector of the ith agent. Suppose that the utility function of the ith agent, i ∈ H, can be written as

\[ u_i(x_i; \alpha, \beta) = -\sum_{j=1}^{n} \frac{\beta_{ij}^{2} + 1}{x_{ij}^{2}}, \]  

where xj ≥ 0, α > 0, β = (β1, β2, . . . , βn), and β1 ≥ 0. This utility function is the n-dimensional version of Scarf’s C.E.S. function. In what follows, let us examine the case: \{ui(x; α, β), wi\} where

\[ \begin{align*}
\alpha &\in (1, \infty), \\
\beta_i &= (\beta_1, \cdots, \beta_{i-1}, \beta_i, \beta_{i+1}, \cdots, \beta_n) \\
&= (0, 0, \beta, 1, 0, \cdots, 0), \\
\beta &\in (0, \infty), \\
\omega_i &= (\omega_1, \cdots, \omega_{i-1}, \omega_i, \omega_{i+1}, \cdots, \omega_n) \\
&= (0, 0, 1, 0, \cdots, 0),
\end{align*} \]

with \( \beta_{n+1} = \beta_n \). Let \( p = (p_1, p_2, \ldots, p_n) \) ≥ 0 be a price vector in which \( p_j \) is the price of the jth commodity. Now we solve the constrained maximum problem of the ith agent: maximize \( u_i(x_i; \alpha, \beta) \) s.t. \( \langle p, x_i \rangle \leq \langle p, w_i \rangle \) and \( x_i \geq 0 \). This yields the demand function for the ith agent, \( x_i(p; \alpha, \beta) = (x_{i1}(p; \alpha, \beta), x_{i2}(p; \alpha, \beta), \ldots, x_{in}(p; \alpha, \beta)) \) where

\[ x_{ij}(p; \alpha, \beta) = \begin{cases} 
\frac{\beta p_i^{\alpha/(1+\alpha)} p_i^{1/(1+\alpha)} \beta p_i^{\alpha/(1+\alpha)} + p_i^{\alpha/(1+\alpha)} - \beta p_i^{\alpha/(1+\alpha)} (1+\alpha)}{\beta p_i^{\alpha/(1+\alpha)} (1+\alpha)}, & \text{if } j = i, \\
\frac{\beta p_i^{\alpha/(1+\alpha)} (1+\alpha)}{p_i^{1/(1+\alpha)} + \beta p_i^{\alpha/(1+\alpha)}}, & \text{if } j = i + 1, \\
0, & \text{otherwise},
\end{cases} \]
with $x_{n+1} = x_n$. Then the market excess demand function for the $j$th commodity, $\zeta_j(p; \alpha, \beta) := \sum_{n} x_{ij}(p; \alpha, \beta) - \omega_j$, is given by

$$
\zeta_j(p; \alpha, \beta) = \frac{\beta p_j^\alpha}{\beta p_j^\alpha + p_{j+1}^{\alpha(1+\alpha)}} + \frac{p_{j+1}}{p_j + \beta p_{j+1}^{\alpha(1+\alpha)}} - 1, \quad j \in H, \tag{2}
$$

with $p_0 = p_n$ and $p_{n+1} = p_1$.

Therefore the economy $\mathcal{E}^n = ((u(x_j; \alpha, \beta), \omega_i)_{i=1}^n)$ has the excess demand function $\zeta(p; \alpha, \beta) := (\zeta_1(p; \alpha, \beta), \zeta_2(p; \alpha, \beta), \ldots, \zeta_n(p; \alpha, \beta))$, and we have $\zeta(p; \alpha, \beta) = (0, 0, \ldots, 0)$ for $\beta = \hat{\beta}$ where $\hat{p}_1 = \hat{p}_2 = \cdots = \hat{p}_n$ and this implies that $\hat{p}$ is a competitive equilibrium of $\mathcal{E}^n$. The excess demand function has the following properties:

- (P1) if $n = 2, \beta = 2, \text{ and } \alpha \to \infty$, it goes to Gale [5]'s that.
- (P2) if $n = 2, \beta = 2, \text{ and } \alpha/(1 + \alpha) = \mu$, it becomes Bala [2]'s that.
- (P3) if $n = 3, \beta = 1, \text{ and } \alpha \to \infty$, it goes to that in Scarf’s first example.
- (P4) if $n = 3$, it becomes that in Scarf’s second example.
- (P5) $\delta \zeta_j(p; \alpha, \beta)/\delta p_j > 0$ and $\delta \zeta_j(p; \alpha, \beta)/\delta p_{j+1} < 0$ for every $j \in H$.

Here (P1) and (P2) may not be well known, (P3) is stated by Scarf, (P4) and (P5) are immediate consequence of the present extension, and Scarf’s ‘first’ example indicates the economy with the fixed proportions function instead of (1).

Then a tâtonnement is given by $dp_j/dt = \zeta_j(p; \alpha, \beta)$, and we consider the normalized system in which the $n$th commodity is taken to be a numéraire and the $n$th equation is dropped because of the Walras law: $\sum_{j=1}^n p_j \delta_j(p; \alpha, \beta) = 0$,

$$
\mathcal{A}^n \frac{dp_j}{dt} = \zeta_j(p; \alpha, \beta), \quad j \in I = H \setminus \{n\},
$$

with $p(0) \gg 0$.

**Lemma 1.** Consider a system of parametric differential equations,

$$
\frac{dz}{dt} = f(z; \epsilon), \quad z \in \mathbb{R}^n, \quad \epsilon \in \mathbb{R},
$$

with a smooth function $f$, which has an equilibrium $(z^*; \epsilon^*)$. We assume that the Jacobian matrix of the system is given by $J(z^*; \epsilon) = (a_j(\epsilon))$ where $a_j(\epsilon) = a(\epsilon) \delta_{ij} + b(\epsilon) \delta_{i+1,j} + c(\epsilon) \delta_{i-j,1}$, ($i, j = 1, \ldots, m$) and $\delta_{ij}$ is the Kronecker delta, $\delta_{ij} = 0$ for $i \neq j$ and $\delta_{ij} = 1$ for $i = j$, that is, the Jacobian matrix has the tridiagonal Toeplitz form

$$
J(z; \epsilon) = \begin{pmatrix}
    a(\epsilon) & b(\epsilon) & 0 \\
    c(\epsilon) & a(\epsilon) & b(\epsilon) \\
    c(\epsilon) & a(\epsilon) & \ddots & \ddots \\
    \vdots & \ddots & \ddots & b(\epsilon) \\
    0 & \cdots & c(\epsilon) & a(\epsilon)
\end{pmatrix}. \tag{3}
$$

We also assume that $a(\epsilon), b(\epsilon),$ and $c(\epsilon)$ are smooth functions of $\epsilon$.

Suppose that

$$
a(\epsilon^*) = 0, \quad b(\epsilon^*)c(\epsilon^*) < 0, \quad \frac{d}{d\epsilon}(a(\epsilon)) \neq 0, \quad (4)
$$

then there is a bifurcation with all eigenvalues passing through the imaginary axis in the same direction. In addition, all the eigenvalues are distinct and, at $\epsilon = \epsilon^*$, purely imaginary but one zero eigenvalue for odd $m$.

**Proof.** The eigenvalues of this class of matrices can be found analytically (see, e.g., Todd [12, pp. 155-156]),

$$
\lambda_k = a(\epsilon) + 2 \sqrt{b(\epsilon)c(\epsilon)} \cos \left( \frac{k\pi}{m+1} \right), \quad k = 1, 2, \ldots, m,
$$

from which the assertion follows. □

**Proposition 1.** Let $p = (p_1, p_2, \ldots, p_{n-1})$ and $\beta^* = (\alpha + 1)/(\alpha - 1)$. For every $\epsilon$, the system $\mathcal{A}^n$ has an equilibrium $p^* = (1, 1, \ldots, 1)$. The equilibrium is locally stable for $\beta \in (0, \beta^*)$ and locally completely unstable (all eigenvalues have positive real parts) for $\beta \in (\beta^*, \infty)$.

**Proof.** It follows immediately from $\hat{p}$ with the normalization that $p^* = (1, 1, \ldots, 1)$ is an equilibrium of the system $\mathcal{A}^n$. Let $\beta$ be a bifurcation parameter. Observing (2) with $p_n = 1$ for $j \in I$, we find that the Jacobian matrix $J(p^*; \beta)$ has the form in (3) where $a(\beta) = -(1 + \alpha + \beta - a\beta)/(1 + \alpha)(1 + \beta)^2$. Then we have $a(\beta^*) = 0$ and $\beta^* = (\alpha + 1)/(\alpha - 1) > 0$ for $\alpha \in (1, \infty)$. The inequalities in (P5) imply $b(\beta^*)c(\beta^*) < 0$. Moreover, we obtain $da(\beta^*)/d\beta = (a - 1)/\{(1 + \alpha)(1 + \beta)^2\}^2 > 0$ for $\alpha \in (1, \infty)$. Therefore the set of conditions (4) of Lemma 1 is fulfilled with the positive crossing velocity. □

Let us note that Scarf did not consider the normalized system in the second example, i.e., $\mathcal{A}^n$ with (P4), and that the structure depends on the normalization. Due to the circularity of the $n$-dimensional version of the Scarf example, the structure does not depend on the choice of the numéraire. For the problem arising from changing the numéraire, see Veendorp [13] for Scarf’s examples and Mukherji [7] for Gale’s example.

Bala [2] extended Gale’s example to a parametric economy and showed a pitchfork bifurcation in the normalized system equivalent to $\mathcal{A}^n$ with (P2), as the parameter $\mu$ crosses a critical value, the unique globally stable equilibrium loses local stability and gives rise to two new locally stable equilibria on either side. The system including Gale’s one has exactly the structure above with $m = 1$. Mukherji [8] extended Scarf’s first example to a parameterized economy with respect to $\omega_{22}$ and presented that the normalized system has a Hopf bifurcation: as the parameter passes a critical value, the unique globally stable equilibrium becomes locally unstable and periodic orbits appear. The system including that in Scarf’s first example does not
have the structure above with \( m = 2 \) due to the asymmetry of endowments, but the qualitative property of the system is similar to \( \mathcal{L} \) with (P4).

Dierker [3, 4] proved by using the properties of the fixed point index, that, given a desirability assumption for price adjustment processes in a regular economy, if there is a uniquely locally completely unstable equilibrium, then the number of commodities is odd. This means that Scarf’s instability result is impossible for even number of commodities (see Dierker [4, Secs. 5, 11]). In fact, the following Proposition 2 is established.

**Proposition 2.** If \( n \) is odd, for all \( \beta \in (0, \infty) \), the system \( \mathcal{L}^n \) has a unique equilibrium \( p^* \). If \( n \) is even, the system \( \mathcal{L}^n \) has a unique equilibrium \( p^* \) for \( \beta \in (0, \beta^*) \) and exactly three equilibria, for \( \beta \in (\beta^*, \infty), p^*, \hat{p}, \tilde{p} \) where \( \hat{p}_j = q \in (0, 1) \) for odd \( j \) and \( \hat{p}_j = 1 \) for even \( j \), and \( \tilde{p}_j = r \in (1, \infty) \) and \( r = 1/q \) for odd \( j \) and \( \tilde{p}_j = 1 \) for even \( j \). Here \( j \in I \).

**Proof.** Recall that \( \hat{p} \) where \( \hat{p}_1 = \hat{p}_2 = \cdots = \hat{p}_n \) is one equilibrium of \( \mathcal{L}^n \). Suppose that there exists an equilibrium \( \hat{p} \neq \hat{p} \). Considering (P5), we have two possible cases for some \( \xi_j(\hat{p}; \alpha, \beta), j \in J \), with (i) \( \hat{p}_j < \hat{p}_{j-1}, \hat{p}_{j+1} \) and with (ii) \( \hat{p}_j > \hat{p}_{j-1}, \hat{p}_{j+1} \), since the other cases violate \( \xi_j(\hat{p}) = 0 \). Now we assume that (i) holds. As we can take \( j = 1 \) without loss of generality, we then have \( \hat{p}_1 < \hat{p}_n, \hat{p}_2 \). Using this and again (P5), we obtain \( \hat{p}_2 > \hat{p}_1, \hat{p}_3 \), and in a similar way we successively have \((\hat{p}_2 < \hat{p}_3, \hat{p}_4), (\hat{p}_3 > \hat{p}_4, \hat{p}_5), \cdots \), which induce \( \hat{p}_n < \hat{p}_{n-1}, \hat{p}_1 \) for odd \( n \) and this is a contradiction but \( \hat{p}_n > \hat{p}_{n-1}, \hat{p}_1 \) for even \( n \) and this is not. For the case of (ii), a similar argument, a contradiction is obtained for odd \( n \) but not for even \( n \). The above results show that if \( n \) is odd, \( \hat{p} \) is a unique equilibrium of \( \mathcal{L}^n \) and then \( p^* \) is a unique equilibrium of \( \mathcal{L}^n \). These also suggest that if \( n \) is even, other equilibria of \( \mathcal{L}^n \) can be possible for \( \hat{p} \) where \( \hat{p}_j \) satisfies (i) for every odd \( j \) \( \in H \) and for \( \tilde{p} \) where \( \tilde{p}_j \) fulfills (ii) for every odd \( j \) \( \in H \).

Suppose that \( n \) is even. Let \( \hat{p} \) be a vector with \( \hat{p}_j \) satisfying (i) for every odd \( j \) \( \in H \) and \( \tilde{p} \) where \( \tilde{p}_j \) fulfills (ii) for every odd \( j \) \( \in H \).

Now we take \( j = 1 \). Since the right hand side is positive, so is the left hand side. This fact with \( \hat{p}_1 \in (0, 1), \hat{p}_n \in (1, \infty) \) implies \( \hat{p}_2 \geq \hat{p}_n = 1 \). Given this, in a similar way for \( \xi_j(\hat{p}; \alpha, \beta) = 0 \), we obtain \( \hat{p}_1 \geq \hat{p}_2 > \hat{p}_n = 1 \). Repeating the procedure for \( \xi_j(\hat{p}; \alpha, \beta) = 0 \) for every odd \( j \) \( \in H \), we successively have \( \hat{p}_{n-2} \geq \cdots > \hat{p}_{n-1} \geq \hat{p}_n = 1 \). Next, conversely, we start from \( j = n - 1 \). By the argument above, we have \( \hat{p}_{n-2} \leq \hat{p}_{n-1} = 1 \), and then successively \( \hat{p}_2 \leq \cdots \leq \hat{p}_{n-2} \leq \hat{p}_n = 1 \). These two relations show \( \tilde{p}_j = 1 \) for every odd \( j \) \( \in H \). Therefore, together with \( \hat{p}_j = q \in (0, 1), \tilde{p}_{j-1} = \hat{p}_{j+1} = 1 \), (5) becomes

\[
\phi(q) := \beta q^{1/(1+\alpha)} - q^\alpha/(1+\alpha) = 1 - q. \tag{6}
\]

Then we have \( \lim_{q \to 0} \phi(q) = 1 = \phi(1) = 0, \phi(q) > 0 \) for all \( q \in (0, 1), \lim_{q \to 1} d\phi(q)/dq = +\infty, \) and \( d^2\phi(q)/dq^2 < 0 \) for all \( q \in (0, 1) \). Moreover, we have \( |d\phi(1)/dq| \geq 1 \iff \beta \geq \beta^* \). These imply that when \( \beta \in (0, \beta^*) \) there is no value \( q \in (0, 1) \) satisfying (6) and when \( \beta \in (\beta^*, \infty) \) there is exactly one value \( q \in (0, 1) \) satisfying (6). From the symmetry of the equation \( \xi_j(\hat{p}; \alpha, \beta) = 0 \) for every odd \( j \) \( \in H \), it follows that odd \( j \) is arbitrary and thus \( q \) is identical for every odd \( j \) \( \in H \). Then, (6) is also derived from the equation \( \xi_j(\hat{p}; \alpha, \beta) = 0 \) where \( \hat{p}_j = 1, \tilde{p}_{j-1} = \hat{p}_{j+1} = q \) for any even \( j \) \( \in H \). Therefore, for \( \beta \in (\beta^*, \infty), \) the vector composed of the first \( n - 1 \) elements of \( \hat{p}, \tilde{p} = (q, 1, \ldots, q) \), is another equilibrium of the system. Substituting \( 1/q \) into (6) and multiplying \(-1\) to each side shows that \( 1/q = r \in (1, \infty) \) also satisfies (6). As \( q \) is unique value in \( (0, 1) \), so is \( r \in (1, \infty) \). The above results show that, for \( \beta \in (\beta^*, \infty), \) the vector replaced \( q \) in \( \hat{p} \) with \( r, \tilde{p} = (r, 1, \ldots, r) \) is the other equilibrium of the system. It is clear that \( \hat{p}_j \) satisfies (ii) for every odd \( j \) \( \in H \).

\[\square\]

The idea of the proof for odd \( n \) is an extension of that of Lemma 2 in Scarf [11]. Proposition 2 in conjunction with Proposition 1 says that Scarf’s instability result holds for the system with odd numbers of commodities in the present example. For a construction of examples like Scarf’s in general settings, see Lemma 5 in Saari and Simon [10]. On the other hand, the idea of the proof for even \( n \) is an extension of that of Lemma 1 in Bala [2]. The Proposition 2 together with Proposition 1 indicates that when the number of commodities is even, Gale’s result takes place.

**Figure 1:** Stability boundary of \( p^* \) in \( \mathcal{L}^n \).

Figure 1 shows the instability (white/gray) region of \( p^* \) in the \((\mu, \beta)\)-plane of the economy, \( \mathcal{L}^n \), for every \( n \). Here \( \alpha/(1 + \alpha) = \mu \). When \( \alpha \to \infty \) (or \( \mu \to 1 \)), the indifference curves become L-shaped, and instability is most likely to occur. The smaller the average of \( \alpha \) (or \( \mu \)), the greater the degree of substitutability between the commodities and the smaller the portion of instability. When \( \alpha \to 1 \) (or \( \mu \to 1/2 \)), instability disappears. This effect is sharpened by Bala [2] and Hirota [6]. On the other hand, when

---

225
$β \to \infty$, each agent places a highest value on his own commodity, and instability is most likely to obtain. The smaller the value of $β$, the less desire of each agent for his own commodity and the smaller the part of instability. When $β < 1$, instability vanishes. This effect is recognized by Gale [5].

Summary 1. The economy $E^n$ has a unique locally stable equilibrium at $p^*$ for $β \in (0, β^*)$ in the normalized system. If $n$ is odd, for $β \in (β^*, \infty)$, the equilibrium becomes locally completely unstable and is still unique. If $n$ is even, for $β \in (β^*, \infty)$, the equilibrium becomes locally completely unstable but two other equilibria appear. In the economy, stability arises when income effects are smaller and each agent is less interested in his own commodity, while instability appears when income effects are larger and each agent is more interested in his own commodity.

3. Conclusion

The results allow us to classify the case with odd numbers of commodity as the Scarf type bifurcating economy and the case with even numbers of commodity as the Gale-Bala type bifurcating economy.

Appendix

References


Figure 2: (Top): Bifurcation diagram showing a supercritical pitchfork bifurcation for $n = 2$ and $α = 2$. (Bottom): Bifurcation diagram showing a supercritical Hopf bifurcation for $n = 3$ and $α = 2$. 

226
Induced Technical Change and Economic Fluctuations: An Imperfect Competition Model

Hiroyuki Yoshida

†Department of Economics, Nihon University
1-3-2 Misaki, Chiyoda-ku, Tokyo 101-8360 JAPAN
Email: yoshida@eco.nihon-u.ac.jp

Abstract—In this paper we develop the Malinvaud (1982) model, which studies the process of economic growth when prices are sticky and production technology is putty-clay. Our model is formulated on the explicit microfoundations of firm’s behavior: firms maximize their profit by using the subjective demand curve and minimize the unit cost of production with new capital equipment. Our main concern is to examine the possibility of endogenous and persistent fluctuations in a dynamic real sector model. In fact, with the aid of the Hopf bifurcation theorem, we show that periodic oscillations occur around the steady state if the Kaldor (1940) condition is satisfied.

1. Introduction

Investment spending is of great importance in macroeconomic theory. In the long run capital accumulation is one of the main determinants of economic growth. What is more, the movement of investment occupies a prominent position in the analysis of business cycles. Since investment is much more volatile than consumption, it is believed that the volatility of investment is a source of output fluctuations.

The purpose of the present paper is to investigate the occurrence of endogenous and perpetual business cycles in the Keynesian framework. In particular, we develop the Malinvaud (1982) model, which studies the process of economic growth when prices are sticky and production technology is putty-clay. Compared to the Malinvaud model, our model is formulated on the explicit microfoundations of firm’s behavior: firms maximize their profit by using the subjective demand curve and minimize the unit cost of production with new capital equipment. In such a setting we shall examine the effects of induced technical change in the process of business cycles.

The rest of paper is organized as follows. Section 2 explains the distinction between the short-run and the long-run production functions. Section 3 sets up the formal model. In Section 4 we investigate the dynamic properties of the model. The final section summarizes our results.

2. Long-Run and Short-Run Production Functions

In this section we shall explain the long-run and the short-run production functions.\(^1\) In the first place, we consider the long-run production function that plays an important role at the time of the choice of the best manufacturing techniques. It is given by

\[ Y_n(t) = F(N_n(t), K(t)), \]

where \(Y_n(t)\) stands for the level of capacity output at time \(t\). Furthermore, we assume that the long-run production function is homogeneous of degree one. We can therefore write in the form,

\[ Y_n(t)/K(t) = f(x_n(t)), \quad x_n(t) = N_n(t)/K(t), \]

where \(f(0) = 0, f(\infty) = \infty, f'(\cdot) > 0, \) and \(f''(\cdot) < 0.\)

In the second place, we shall explain the short-run production function that governs the current production decisions with the existing capital stock. It is usual that the level of aggregate demand does not agree with the level of capacity output in our free market economy. In such a situation firms match the actual level of output to the level of aggregate demand by controlling the degree of capital utilization. If aggregate demand exceeds potential output, firms need the amount of labor that is greater than \(N_n\) to overutilize the capital stock. In the opposite case, firms employ the amount of labor that is less than \(N_n\) to underutilize the capital stock. The above relationship can be represented by

\[ u = u(N/N_n), \quad u'(\cdot) > 0, u(0) = 0, u(1) = 1, \]

where \(u\) represents the rate of capacity utilization. Using (3) we can therefore express the short-run production function as follows:

\[ \frac{Y}{K} = \frac{Y_n}{K} = u(x/x_n)f(x_n), \quad x = N/K, \]

where \(Y\) is the actual level of output.

Now we are in a position to point out the envelope property of the long-run production function. As is well known in microeconomic theory, the long-run cost function is the lower envelope of the family of the short-run cost functions. Such a relationship exists also in our framework:

\(^1\)For further details of this subject, see Yoshida (1999).
the long-run production function is the lower envelope of the family of the short-run production functions. Thus, we must have

\[ f'(x_n) = u'(1 - \frac{f(x_n)}{x_n}) \quad \text{for all } x_n. \]  

(5)

This condition implies that the long-run and the short-run production functions are tangent to each other at any normal employment-capital ratio \( x_n \). By integrating (5), we obtain

\[ f(x_n) = A(x_n)^\theta \quad A > 0, \quad 0 < \theta < 1, \]  

(6)

which is the Cobb-Douglas production function itself.

For the moment, we shall confine our attention to the evolution of technology. As mentioned in Introduction, we adopt the assumption of putty-clay technology. Recalling the definition of \( x_n \), we have

\[ x_n(t) = \frac{N_\theta(t)}{K(t)} = \frac{\int_0^t x_n^{(v)}(t) e^{-\delta(t-v)} I(v) dv}{\int_0^t e^{-\delta(t-v)} I(v) dv}, \]  

(7)

where \( x_n^{(v)}(t) \) is the labor intensity required to operate equipment of vintage \( v \) at full capacity, \( \delta \) is the capital depreciation rate, and \( I(v) \) is the amount of newly installed equipment at time \( v \). Taking the logarithmic derivative of (7) with respect to time leads to

\[ x_n = g(x_n^{(0)}(t) - x_n(t)), \quad g = I/K \]  

(8)

where \( g \) represents the rate of capital accumulation. This equation implies that the labor intensity of the capital stock evolves not instantaneously but at the speed of \( g \).

Before closing this section, we add a remark on the treatment of the utilization rate. Throughout this paper we shall emphasize the role of \( \pi \) in the process of business cycles. Hence, we take the inverse function of (3) to explicitly consider the utilization rate:

\[ N/N_n = G(u), \]  

(9)

with \( G(0) = 0, \ G(1) = 1, \ G'(1) = f(x_n)/(x_n f'(x_n)) = 1/\theta. \)

3. The Model

We construct a simple model with imperfect competition in the output market. There are two types of agents in the economy: firms and households. The output market is characterized by imperfect competition. The representative firm does not have full and perfect information about the “true” demand curve facing her, and has the following iso-elastic demand function

\[ Y = B p^{-1/\varepsilon} \quad 0 < \varepsilon < 1, \]  

(10)

where \( p \) is the price level, \( B \) is a scale parameter that reflects the expected level of demand, and \( \varepsilon \) is a constant parameter that represents the price elasticity of demand. Note that the parameter \( B \) is a function of time.

Using (4), (9), and (10), the operating profit of the firm, \( \pi \), is defined as follows

\[ \pi = pY - WN, \]  

(11)

\[ = [(uf(x_n)k)^{-\varepsilon} uf(x_n) - WG(u)x_n]K, \]

where \( k = K/R \), and \( W \) is the nominal wage rate.

The firm maximizes her profit \( \pi \) by controlling the rate of capital utilization. The optimal condition for the firm is obtained as follows,

\[ (1 - \varepsilon)[uf(x_n)k]^{-\varepsilon} f(x) = WG'(u)x. \]  

(12)

To assure the second-order condition of profit maximization, we must add the following assumption:

**Assumption 1:**

\[ \varepsilon + uG''/G' > 0. \]  

(13)

This condition does indeed lead to the negative sign of the second-order condition with respect to \( u \):

\[ \frac{d^2 \pi}{du^2} = -\frac{WG'x_n}{u} (\varepsilon + uG''/G') < 0. \]  

(14)

By applying the implicit function theorem, we can transform (12) into the following equation

\[ u = u(k, x_n), \]  

(15)

where

\[ u_k = \frac{\partial u}{\partial k} = \frac{u}{k} \frac{\varepsilon}{k + uG''(u)/G'(u)} < 0, \]  

(16)

and

\[ u_x = \frac{\partial u}{\partial x_n} = -\frac{u}{x} \frac{1 - \varepsilon(1 - \varepsilon)}{x \varepsilon + uG''(u)/G'(u)} < 0. \]  

(17)

Then we shall turn to the issue of technical change. As mentioned above, in our framework new production techniques are embodied in new equipment. The firm chooses \( x_n^{(0)}(t) \) to minimize the unit cost of production with new capital equipment at time \( t \). The objective function for this problem is:

\[ \frac{wN + rI}{uF(N, I)} = \frac{wG(u)x_n + r}{uf(x_n)}, \]  

(18)

where \( w \) is the real wage rate and \( r \) is the real rate of interest. Let us assume that the individual firm is a price taker, i.e., she takes \( w \) and \( r \) as given, exogenous variables to the cost-minimizing problem. Thus, the first-order condition takes the following form:

\[ \frac{f(x_n) - x_n f'(x_n)}{f'(x_n)} = \frac{r}{wG(u)}. \]  

(19)

From (6), (12), and (19), some straightforward manipulations yield

\[ x_n = \frac{\theta}{1 - \theta} \frac{r}{1 - \varepsilon} \frac{G'(u)(x_n)^{1-\theta}}{A} = \phi(x_n, u), \]  

(20)

\[ ^2\text{Here and henceforth, we omit the vintage index so long as no confusion occurs.} \]
where
\[
\frac{\partial x_e}{\partial x_n} = \phi_x = (1 - \theta) x_n > 0, \tag{21}
\]
\[
\frac{\partial x_e}{\partial u} = \phi_u = \left( \frac{uG''(u)}{G'(u)} - \frac{uG'(u)}{G(u)} \right) x_u = \Phi(u) x_u. \tag{22}
\]

We assume that the investment behavior is independent of the saving behavior on the line of the Keynesian tradition. To begin with, it is assumed that the accumulation rate, \( g = I/K \), depends on the rate of capacity utilization:
\[
g = g(u), \quad g' > 0. \tag{23}
\]

Next, we turn to the saving function. Let saving \( S \) be a linear function of income:
\[
S = sY, \quad 0 < s < 1, \tag{24}
\]
where \( s \) is the marginal propensity to save.

The output market is usually in a state of disequilibrium. Thus the firm constantly adjust the scale parameter \( B \) which reflects the expected level of demand. The evolution of \( B \) will be formulated by
\[
\frac{B}{B} = \omega \frac{I - S}{K} + \mu, \quad \omega > 0, \tag{25}
\]
which says that the firm increases \( B \) when aggregate demand exceeds output (\( I > S \)) and vice versa. Note that there is a trend term \( \mu \) in the growth rate of \( B \). For simplicity we assume that \( \mu \) is the steady-state growth rate of output:
\[
\mu = (\dot{Y}/Y)' \tag{26}
\]

4. Local Stability and the Hopf Bifurcation

The dynamic system is now formed by the following two differential equations:
\[
k = [g(u(k, x_n)) - \delta - \mu - \omega(g - s)(k, x_n)f(x_n)]k, \tag{27}
\]
\[
x_n = g(u(k, x_n))(\phi(u(k, x_n), x_n) - x_n). \tag{28}
\]

The non-zero steady-state solution is given by \((k^*, x_n^*)\) such that
\[
g(u(k, x_n)) - \delta - \mu = \omega g - s(k, x_n)f(x_n), \tag{29}
\]
\[
\phi(u(k, x_n), x_n) = x_n. \tag{30}
\]

Notice that \( \mu = (\dot{Y}/Y)' = g' - \delta \) in the steady state, where \( g' \) is the steady-state rate of accumulation. Thus it follows from (29) that investment and saving are equal in the steady state:
\[
g(u(k^*, x_n^*)) = s(k^*, x_n^*)f(x_n^*). \tag{31}
\]

To study the local stability of the dynamic system, we have to investigate the Jacobian matrix \( J \) evaluated at the steady-state point. We obtain
\[
J = \begin{bmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22}
\end{bmatrix}, \tag{32}
\]
where \( a_{11} = [g'u_s - \omega(g' - s)f_u]k, a_{12} = [g'u_s - \omega g'u_s - s(u_s f + u'_f)]k, a_{21} = g\phi_u u, \) and \( a_{22} = g(\phi_u + \phi_x - 1) \).

The corresponding characteristic equation can be written as
\[
\lambda^2 + b_1 \lambda + b_2 = 0, \tag{33}
\]
where
\[
b_1 = (g' - s)u_k \omega - g'u_k - g(\phi_u - \theta), \tag{34}
\]
\[
b_2 = [\theta(g' - s) - \phi_u s f'(u_k)g\omega - g'u_kg\theta]. \tag{35}
\]

To incorporate a positive feedback effect in the output adjustment mechanism, we assume

Assumption 2:
\[
g'(u') > sf. \tag{36}
\]

This condition is the same as one of Kaldor’s (1940) conditions. He emphasized the importance of this condition to establish the existence of limit cycles that fluctuate around the steady state. We furthermore adopt the following assumption:

Assumption 3:
\[
\phi_u > \theta(g' - s)/(sf'). \tag{37}
\]

If we choose \( \omega \) as a bifurcation parameter, we can see that \( b_1 = b_1(\omega) \) is a monotonically decreasing function with a positive vertical intercept and that \( b_2(\omega) > 0 \) for all \( \omega > 0 \). We can now prove the following proposition:

Proposition:
(1) There exists a critical value \( \omega_H > 0 \) such that the steady state of the system is locally stable for \( \omega < \omega_H \) and unstable for \( \omega > \omega_H \), where \( \omega_H \) is the unique solution of \( b_1(\omega) = 0 \).

(2) The steady-state point loses stability in a cyclical manner at \( \omega = \omega_H \) by way of the Hopf bifurcation. We can therefore observe the endogenous and persistent business cycles that oscillate around the steady state.

5. Conclusion

In this paper we extended the Malinvaud (1982) model, and showed the emergence of endogenous and self-sustained fluctuations.

The argument for the occurrence of cyclical fluctuations runs as follows. Suppose that some shock increases \( u \) at the steady state. This rise in \( u \) stimulates capital accumulation, and hence causes further expansion (\( g' > sf \)). This effect is called the Kaldor effect, which is a destabilizing force. In contrast, we have a stabilizing force, which results from induced technical change. The increase in \( u \) also makes the firm choose more labor-using directions of technical change (\( \phi_u > 0 \)). This tendency eventually induces an increase in \( x_n \). Consequently, the above process causes
a reduction in $u (u_x < 0)$, which leads to a slowdown in expansion. In this manner, the interaction of the two forces yields the cyclical movements in our dynamic model.

References


Simple Business Cycle Model with Policy Lags

Akio Matsumoto
Department of Economics
Chuo University
742-1, Higashi-Nakano, Hachiji
Tokyo, 742-1, Japan
akiom@tamacc.chuo-u.ac.jp

Mami Suzuki
Department of Management
Aichigakusen University
1 Shiotori, Ohikemachi
Toyoda, Aichi, 471-8532, Japan
m-suzuki@gakusen.ac.jp

Abstract

This paper studies how a time lag in policy response affects the stability of income adjustment process. Two different approaches are taken: one is local stability analysis and the other is to consider the dynamical system with continuously distributed lags. It is demonstrated that the time-lag has destabilizing effects. It is also demonstrated that the length of time-lag and the strength of the policy response can be sources of instability.

1 Basic Model

We recapitulate the basic elements of Kaldor’s Model and add our specification of policy lag. We denote national income by $Y$, investment by $I$, savings by $S$, Government expenditure by $G$ and $K$ by the capital stock. The model can be summarized in the following equations:

$$\dot{Y}(t) = \alpha \{I(Y(t),K(t)) - S(Y(t)) - G(t)\},$$

$$\dot{K}(t) = I(Y(t),K(t)) - \delta K(t).$$

where the dot over a variable means a time derivative, $\alpha$ is an adjustment coefficient, and $\delta$ is the constant depreciation rate. The first equation states that income changes proportionally to the excess demand in the good market. The second states that the accumulation or decumulation of the capital stock. The main aim of this study is to find how the policy-lag affects dynamics of output. To this end, we simplify the model by making the following three assumptions:

Assumption 1. $K(t)$ is constant for all $t \geq 0$.

Assumption 2. $I(Y) = I_Y Y$ with $I_Y > 0$, $S(Y) = S_Y Y$ with $S_Y > 0$, and $G(t) = \beta Y(t - \tau)$ with $\beta > 0$.

Assumption 3. $a = I_Y - S_Y > 0$.

Assumption 1 reduces Kaldor’s model to a one dimensional dynamic model in which output $Y(t)$ is a time variable. Assumption 2 confines the following analysis to the case in which behavioral specifications are linear in income and only the lag in policy response is taken into account. Dependency of investment on capital is implicitly assumed away due to Assumption 1. Assumption 3 is Kaldor’s basic condition for the normal level of income.

As a benchmark, we start to consider the model without the time lag (i.e., $\tau = 0$). Under Assumptions 1, 2 and 3, the dynamical system is reduced to

$$\dot{Y}(t) = \alpha(a - \beta)Y(t).$$

This is a one-dimensional linear differential equation and it can be seen that a stationary state is zero and unique. Its solution is given by

$$Y(t) = Y_0 e^{\alpha(a - \beta)t},$$

and is stable or unstable according to whether $a - \beta$ is negative or positive. Since we are interested in how time lags disturb an otherwise stable dynamical system, we make the following assumption for the stability of the model without time lag.

Assumption 4. $a - \beta < 0$. 

231
2 Local Stability with Time Lags

We explicitly introduce a time lag in policy response and show that such a time lag can be a source of cyclic behavior. The dynamical system with policy lag is written as

$$\dot{Y}(t) = \alpha \{aY(t) - \beta Y(t - \tau) \}.$$  

As usual we suppose that the system has an exponential part and solving the result for the characteristic equation, we have

$$\lambda - \alpha a + \alpha \beta e^{-\lambda \tau} = 0.$$  

Setting \(\lambda = 0\) reduce the characteristic equation to \(-\alpha + \beta\) that is positive due to Assumption 4. This means that \(\lambda = 0\) cannot be a solution of the characteristic equation and a stability switch (or a cross of the imaginary axis) necessarily occurs with \(\lambda = \pm i \omega\) with \(\omega > 0\). Without loss of generality, we assume that \(\lambda = i \omega\) with \(\omega > 0\) is a root of the characteristic equation for some \(\tau > 0\).

Substituting this solution into the characteristic equation divides the complex root into the following forms of the real part and the imaginary part

$$\omega^2 = \alpha^2 (\beta^2 - a^2)$$

in which \(\beta^2 - a^2 > 0\) due to Assumption 4. Hence

$$\omega = \alpha \sqrt{\beta^2 - a^2} > 0.$$  

Putting \(\theta = \omega \tau\), dividing the real part by the imaginary part and solving the result for \(\theta\) give

$$\theta = \cot^{-1} \left( \frac{a}{\sqrt{\beta^2 - a^2}} \right) \quad \text{for} \quad 0 < \theta < \pi.$$  

We denote the critical value of time lag by \(\tau_e\),

$$\tau_e = \frac{\theta}{\omega} = \cot^{-1} \left( \frac{a}{\alpha \sqrt{\beta^2 - a^2}} \right).$$  

To check the dependency of \(\lambda\) on \(\tau\), we substitute \(\lambda(\tau)\) into the characteristic equation and differentiate it with respect to \(\tau\) to obtain

$$(1 - \alpha \beta e^{-\lambda \tau}) \frac{d \lambda}{d \tau} = \alpha \beta e^{-\lambda \tau}$$

where from the characteristic equation, we have

$$e^{-\lambda \tau} = -\frac{\lambda - \alpha a}{\alpha \beta}.$$  

For convenience, we study \((d \lambda / d \tau)^{-1}\) instead of \(d \lambda / d \tau\).

Thus

$$\text{sign} \left[ \frac{d \lambda}{d \tau} \right] = \text{sign} \left[ \frac{1}{\omega^2 + \alpha^2} \right].$$

Therefore we have

$$\frac{d \lambda}{d \tau} > 0 \quad \text{for} \quad \lambda = i \omega.$$  

This inequality implies that all the roots that cross the imaginary axis at \(i \omega\) cross from left to right as \(\tau\) increases. We summarize the result:

**Theorem 1** The stationary state is uniformly asymptotically stable for \(\tau < \tau_e\), loses it stability for \(\tau = \tau_e\), and bifurcates to a limit cycle for \(\tau > \tau_e\) where \(\tau_e\) is defined by

$$\tau_e = \frac{\arccot \left( \frac{a}{\alpha \sqrt{\beta^2 - a^2}} \right)}{\alpha \sqrt{\beta^2 - a^2}}.$$  

3 Continuously Distributed Time Lags

We consider the dynamical system with continuously distributed time lags. If the government expenditure is based on an expected income, \(Y^c\), and the expected income is a weighted average of past income, the dynamical system can be written as the system of integro-differential equation,

$$\dot{Y}(t) = \alpha \{aY(t) - \beta Y^c(t) \}$$

$$Y^c(t) = \int_0^t w(t-s, T, m) Y(s) ds,$$
where the weighting function, abbreviated by \( w \), is as follows
\[
\begin{align*}
    w &= \begin{cases} \\
    \frac{1}{T}e^{-\frac{t-s}{T}} & \text{if } m = 0, \\
    \frac{1}{m!} \left( \frac{m}{T} \right)^{m+1} (t-s)^m e^{-\frac{m(t-s)}{T}} & \text{if } m \geq 1.
    \end{cases}
\end{align*}
\]

Here \( m \) is a nonnegative integer and \( T \) is a positive real parameter. In the following analysis, to see how the distribution specification affects the stability of the system, we consider two cases: one is the case with \( m = 0 \) for which weights are exponentially declining with most weight given to the most current data and the other is the case with \( m = 1 \) for which zero weight is given to the most current data, rising to maximum at \( s = t - T \) and then declining exponentially thereafter.

### 3.1 \( m = 0 \): Exponentially Declining Weights

Differentiating \( Y^e(t) \) with respect to time yields the expectation adjustment equation in which expected income is revised according to the difference \( Y(t) \) over \( Y^e(t) \). With this, we can transform the 1D system of the integro-differential equation to a 2D system of ordinary differential equations,
\[
\begin{align*}
    \begin{pmatrix} \dot{Y}(t) \\ \dot{Y}^e(t) \end{pmatrix} &= \begin{pmatrix} \alpha a & -\alpha \beta \\ 1/T & -1/T \end{pmatrix} \begin{pmatrix} Y(t) \\ Y^e(t) \end{pmatrix}.
\end{align*}
\]

The characteristic equation is
\[
\lambda^2 - \left( \alpha a - \frac{1}{T} \right) \lambda + \frac{\alpha}{T} (\beta - a) = 0.
\]

A Hopf bifurcation occurs if the complex conjugate roots cross the imaginary axis. Apparently, the roots are pure imaginary if \( \alpha a - \frac{1}{T} = 0 \). As there are no other real roots in this two-dimensional system, the consideration of the existence of closed orbits is complete if the eigenvalues cross the imaginary axis with nonzero speed at the bifurcation point. We choose \( T \) as the bifurcation parameter. It can directly be seen that there exist a value \( T = \tau_0 \) for which the real part becomes zero,
\[
\tau_0 = \frac{1}{\alpha a} > 0.
\]

The eigenvalues are
\[
\lambda_{1,2} = \pm i \alpha \sqrt{(\beta - a)T}
\]

It is also checked that
\[
\frac{d \text{Re}(\lambda_{1,2})}{dT} = \frac{1}{(\tau_0)^2} > 0.
\]

The inequality implies that the real part crosses the imaginary axis from left to right. Thus the system undergoes a Hopf bifurcation for \( T = \tau_0 \). We summarize the result:

**Theorem 2** When a policy lag on government expenditure has an exponentially declining weight distribution, it might destabilize the otherwise stable one dimensional model if \( T > \tau_0 \) where \( \tau_0 \) is defined by
\[
\tau_0 = \frac{1}{\alpha a}.
\]

### 3.2 \( m = 1 \): Bell-shaped Weight Distribution

When \( m = 1 \), the weighting function is
\[
Y^e(t) = \int_0^t \left( \frac{1}{T} \right)^2 (t-s) e^{-\frac{s}{T}} Y(s) ds.
\]

Introduce a new state variable,
\[
Z(t) = \int_0^t \frac{1}{T} e^{-\frac{s}{T}} Y(s) ds.
\]

Differentiating \( Y^e(t) \) and \( Z(t) \) and replacing the resultant integrals with the new state variable, we can eliminate the integral terms from the dynamic system and, in particular, have the three dimensional system of the ordinary difference equations,
\[
\begin{align*}
    \dot{Y}(t) &= \alpha \{ aY(t) - \beta Y^e(t) \}, \\
    \dot{Y}^e(t) &= \frac{1}{T} \{ Z(t) - Y^e(t) \}, \\
    \dot{Z}(t) &= \frac{1}{T} \{ Y(t) - Z(t) \}.
\end{align*}
\]

The corresponding characteristic equation is
\[
\lambda^3 + a_1(T)\lambda^2 + a_2(T)\lambda + a_3(T) = 0
\]
where coefficients are defined as follows, 

\[ a_1(T) = \frac{2}{T} - \alpha a > 0 \text{ for } T < \frac{2}{\alpha a} \equiv T_1, \]
\[ a_2(T) = \left( \frac{1}{T} - 2\alpha a \right) \frac{1}{T} > 0 \text{ for } T < \frac{1}{2\alpha a} \equiv T_2, \]
\[ a_3(T) = \alpha \frac{T}{T_2} (\beta - a) > 0 \text{ due to Assumption 4.} \]

Since it is seen that \( a_1(T) > 0 \) and \( a_3(T) > 0 \) for \( T < T_2 \), and \( a_3(T) > 0 \) is always true, the roots of the characteristic equation have negative real parts, according to Routh-Hurwitz condition, if and only if \( a_1(T) a_2(T) - a_3(T) > 0 \), namely,

\[ \frac{1}{T^3} \left\{ 2(\alpha a)^2 T^2 - \alpha(4a + \beta) T + 2 \right\} > 0. \]

Let us denote the expression in the parenthesis by \( f(T) \). Then solving \( f(T) = 0 \) for \( T \) gives

\[ T_\pm = \frac{1}{2\alpha a} \left( 4a + \beta \pm \sqrt{3(8a + \beta)} \right). \]

It can be verified that

\[ 0 < T_- < T_2 < T_+. \]

Inequalities imply that

\[ a_1(T) a_2(T) - a_3(T) \geq 0 \text{ according to } T \lesssim T_- \]

Since \( a_1(T) a_2(T) = a_3(T) \) for \( T = T_- \), the characteristic equation can be factorized as

\[ \lambda^3 + a_1(T_-) \lambda^2 + a_2(T_-) \lambda + a_3(T_-) \]

\[ = \left( \lambda^2 + a_2(T_-) \right) \left( \lambda + a_1(T_-) \right). \]

Equating the right hand side to zero and solving it for \( \lambda \) reveals that one of the characteristic root is negative and the other two are pure imaginary,

\[ \lambda_1 = -a_1(T_-) < 0, \]
\[ \lambda_{2,3} = \pm i \sqrt{a_2(T_-)} = \pm i \omega_1. \]

To apply the Hopf bifurcation theorem, we need to determine the sign of the derivative of the real part of the complex roots. We can think of the roots of the characteristic equation as continuous function in terms of the policy lag, \( T \). By implicit differentiation, we have

\[ (3\lambda^2 + 2a_1 \lambda + a_2) \frac{d\lambda}{dT} + (a_1' \lambda^2 + a_2' \lambda + a_3') = 0 \]

implying that

\[ \frac{d\lambda}{dT} = \frac{-a_1' \lambda^2 + a_2' \lambda + a_3'}{3\lambda^2 + 2a_1 \lambda + a_2} \]

where \( \lambda = i \omega_1 \) and \( \omega_1^2 = a_2 \). Rationalizing the right hand side and noticing that the terms with \( \lambda \) are imaginary and the constant and quadratic terms are real yields the following from of the real part of the derivative of \( \lambda \),

\[ \text{Re} \left( \frac{d\lambda}{dT} \right) = - \frac{-a_1' \omega_1^2 + a_2' \left( -3\omega_1^2 + a_2 \right) + 2a_1 a_2' \omega_1^2}{(-3\omega_1^2 + a_2)^2 + (2a_1 \omega_1)^2}. \]

Therefore

\[ \text{sign} \left( \text{Re} \left( \frac{d\lambda}{dT} \right) \right) = - \text{sign} \left( \frac{2(2a_1 T_- - 1)(1 - (\alpha a T_-)^2)}{(T_-)^3} \right), \]

where \( T_- < T_2 \) implies that \( 2a_1 T_- - 1 < 0 \) and \( (\alpha a T_-)^2 < \frac{1}{4} \). It is seen that the sign is positive for \( T = T_- \),

\[ \text{Re} \left( \frac{d\lambda}{dT} \right) > 0. \]

The crossing of the imaginary axis is from left to the right as \( T \) increases and a Hopf bifurcation occurs for \( T = T_- \).

Let \( T_1 = T_- \). The equilibrium is stable for \( T < T_1 \) as Routh-Hurwitz stability conditions are fulfilled, namely, \( a_1(T) > 0, a_2(T) > 0, a_3(T) > 0 \) and \( a_1(T) a_2(T) - a_3(T) > 0 \) while it becomes unstable for \( T \in (T_-, T_2) \). Thus it can be said that a policy lag might have a destabilizing effect, which we sum up as follows.

**Theorem 3** When a policy lag on government expenditure has a bell-shaped weight distribution, it might destabilize the otherwise stable Kaldorian one dimensional model if \( T > T_1 \) where \( T_1 \) is defined by

\[ T_1 = \frac{1}{2\alpha a} \left( 4a + \beta \right) - \frac{\sqrt{3(8a + \beta)}}{2a}. \]
Detection and Quantification of Directional Coupling Using Recurrences

M. Carmen Romano, Marco Thiel and Jürgen Kurths

Institute of Physics, University of Potsdam
Am Neuen Palais 10, 14469 Potsdam, Germany
Email: romano@agnd.uni-potsdam.de, thiel@agnd.uni-potsdam.de, jkurths@agnd.uni-potsdam.de

Abstract—We propose a new index for the detection and quantification of directional coupling in complex systems. This index is applied to weakly as well as strongly coupled systems. This method is based on the recurrence properties of the coupled systems. We investigate the sensitivity of the technique in various numerical experiments and the influence of observational noise.

1. Introduction

The interplay of complex dynamical systems is a central issue in nonlinear dynamics as well as in nonlinear time series analysis. Under certain assumptions different types of synchronization can occur between the interacting systems. This topic has been intensively studied in the last years and has been observed in various fields, such as physics and biology [1]. In such systems it is important not only to analyze the synchronization but also to identify causal (driver-response) relationships. There are mainly three approaches to address this problem: state-space based methods [2], information theory based methods [3] and methods based on the interrelations between the phases of the systems under consideration [4].

In the state-space based approach the state vectors are usually reconstructed by means of delay embedding [5]. If there exists a functional relationship between the driver \( X \) and the response system \( Y \), i.e. \( Y(t) = \Psi(\vec{x}(t)) \), they are said to be generalized synchronized. From the existence of \( \Psi \) it follows that close states of the driver will be mapped to close states of the response. However, if \( \Psi \) is bijective, also close states of the response will be mapped to close states of the driver. Therefore, if \( X \) and \( Y \) are generalized synchronized it is impossible in general to assess the direction of the coupling reliably [6].

Moreover, there are several methods based on information theory to determine the direction of the coupling [3]. They are usually applied to systems which are strongly coupled. In order to treat also weakly coupled systems, the phases of the signals are determined beforehand, and then information theory based indices are applied to the phases. Furthermore, in [4] a technique based on the fitting of the functional relationship between the phases of the two interacting systems has been proposed to detect and quantify the asymmetry in the coupling.

Smirnov and Andrzejak have compared systematically the phase-dynamics approach with the state-space one in the case of weak directional coupling [6]. They have come to the conclusion that none of both approaches is generally superior to the other.

In this paper we introduce a new method to uncover directional coupling. This approach is based on the recurrence properties of both interacting systems. It is rather straightforward to compute, in contrast to the more complicated information theory approaches. Furthermore, it has the advantage that it is applicable to both weak and strong directional coupling.

The outline of this paper is as follows: in Sec. 2 we introduce the measures for the analysis of the directional coupling based on recurrences. In Sec. 3 we demonstrate the proposed measures in three numerical examples. In Sec. 4 we discuss the dependence of the proposed measures on observational noise and finally, we give some conclusions.

2. Recurrence Based Method

Recurrence is a fundamental property of dynamical systems. The concept of recurrence was introduced by Poincaré [7], where he showed that a bounded dynamical system recurs infinitely many times to some neighborhood of a former visited state in phase space. There are many different techniques in nonlinear dynamics which exploit the concept of recurrence. We concentrate of the method of Recurrence Plots (RPs), introduced by Eckmann et al. to visualise the behavior of dynamical systems in phase space [8]. They are defined by means of the recurrence matrix

\[
R_{i,j} = \Theta\left(\varepsilon - ||\vec{x}_i - \vec{x}_j||\right), \quad i, j = 1, \ldots, N, \quad (1)
\]

where \( \vec{x}_i \) denotes the state of the system \( X \) at time \( i\Delta t \) with \( \Delta t \) being the sampling rate, \( \varepsilon \) is a predefined threshold, \( \Theta(\cdot) \) is the Heaviside function and \( N \) is the length of the trajectory considered. The RP is obtained plotting a black dot at the coordinates \((i, j)\) if \( R_{i,j} = 1 \). Looking at the patterns of the RP, one gets a visual impression about the dynamics of the system under consideration. In order to go beyond the visual impression, several measures have been proposed to quantify the patterns in the RP. They have found numerous applications in very different kinds of systems [9]. Moreover, somehow more formal relationships between the patterns obtained in RPs and dynamical invariants have been developed [10]. It has also been shown that the RP contains all topological information about the underlying system [11].
The method of RPs has been extended to Joint Recurrence Plots (JRPs) to analyze the interplay of two or several dynamical systems [12]. The JRP of $X$ and $Y$ is defined as follows
\[ J_{i,j}^{XY} = \Theta(e - ||\vec{x}_i - \vec{x}_j||) \Theta(e - ||\vec{y}_i - \vec{y}_j||), \]
where $\Theta$ is the Heaviside step function. For $CR_{X|Y}$, we consider the mean conditional probability of recurrence (MCR)
\[ CR_{Y|X} = \frac{1}{N} \sum_{i=1}^{N} p(\vec{y}_i | \vec{x}_i) = \frac{1}{N} \sum_{i=1}^{N} \frac{\sum_{j=1}^{N} J_{i,j}^{XY}}{\sum_{j=1}^{N} R_{i,j}}, \]
for the driving system $X$, and
\[ y'_1 = 1.4 - (\mu x_1 y_1 + (1 - \varepsilon) y_1^3) + b_2 y_2, \]
\[ y'_2 = y_1 \]
for the response system $Y$. We use identical systems $b_1 = b_2 = 0.3$. For 101 values of the coupling strength $\mu$ we iterate Eqs. 5,6 and compute an estimate of the correlation dimension $D_2$ [5] for the system $X$ and for the system $Y$ (see Fig. 1). We observe that for $\mu > 0$
the estimate of the correlation dimension $\hat{D}_2(Y)$ for the system $Y$ increases very rapidly. It reaches a maximum at approximately $\mu = 0.09$ and then it decreases more slowly. The onset to identical synchronization occurs at approximately $\mu = 0.65$ [3]. Hence for $\mu > 0.65$ both $\hat{D}_2(Y)$ and $\hat{D}_2(X)$ coincide.

Note that the detection of the directionality holds only before the onset of synchronization. In the case of identical synchronization the series $\{x_i\}$ and $\{y_i\}$ are identical and hence there is no possibility of establishing the causal relationship between $X$ and $Y$ just from the data. This argument can be also extended to the case of generalized synchronization, where the systems are related by a one-to-one function.

3. Numerical Examples

The first example we consider is identical unidirectionally coupled Hénon maps (Eqs. 5,6) with $b_1 = b_2 = 0.3$. We use 10,000 data points for the computation, $\varepsilon = 0.05$, embedding dimension $m = 2$ and delay $\tau = 1$. Analogously to the estimation of the correlation dimension in Sec. 2,
we iterate the maps for 101 values of the coupling strength \( \mu \) and compute the directionality parameters \( CR(X|Y) \) and \( CR(Y|X) \) (see Fig. 2). We observe that \( CR(X|Y) < CR(Y|X) \)

for \( \mu \) between 0 and 0.65 (before the onset to synchronization), indicating correctly the direction of the coupling.

Now we consider two nonidentical unidirectionally coupled Hénon maps (Eqs. 5,6) with \( b = 0.1 \) and \( b = 0.3 \). The embedding dimension in this case is \( m = 3 \) and the delay is \( \tau = 1 \). The onset to generalized synchronization occurs at \( \mu = 0.38 \) [3]. In Fig. 3 we observe that \( CR(X|Y) < CR(Y|X) \)

for \( \mu < 0.38 \), again indicating correctly the direction of the coupling between \( X \) and \( Y \). After the onset to generalized synchronization, the direction of the coupling cannot be determined just from the time series, as argued in Sec. 2. The intersections of both curves \( CR(X|Y) \) and \( CR(Y|X) \) in Fig. 3 can be explained by the dependence of the two largest Lyapunov exponents on the coupling strength \( \mu \) [3].

The last two cases are systems which are strongly coupled. We have seen that the conditional recurrence based method is able to detect correctly the direction of the coupling in these cases. Now we apply the mean conditional probability of recurrence to two systems which are weakly coupled.

Our third example is given by unidirectionally coupled Rössler dynamics. The driving system \( X \) is given by

\[
\begin{align*}
\dot{x}_1 &= -\omega_x x_2 - x_3, \\
\dot{x}_2 &= \omega_x x_1 + 0.15x_2, \\
\dot{x}_3 &= (x_1 - 10)x_3 + 0.2,
\end{align*}
\]  

(7)

and the response system \( Y \) by

\[
\begin{align*}
\dot{y}_1 &= -\omega_y y_2 - y_3 + \mu(x_1 - y_1), \\
\dot{y}_2 &= \omega_y y_1 + 0.15y_2, \\
\dot{y}_3 &= (y_1 - 10)y_3 + 0.2.
\end{align*}
\]  

(8)

The frequency mismatch \( \nu \) is defined by \( \omega_{x,y} = 1 \pm \nu, \) with \( \nu = 0.02 \). We use fourth-order Runge-Kutta routine with a step size of 0.05 and a sampling interval of \( \Delta t = 0.3 \). The length of the time series is 10,000. Time series of \( x_1 \) and \( y_1 \) are taken as observables. The embedding dimension for \( X \) is \( m = 4 \) and the delay \( \tau = 5 \). For the system \( Y \) the embedding dimension is \( m = 5 \) and the delay \( \tau = 5 \). In

![Figure 4: CR(X|Y) (red) and CR(Y|X) (green) for two weakly unidirectionally coupled Rössler systems (Eqs. 7,8) in dependence on the coupling strength \( \mu \).](image)

Fig. 4 we see that \( CR(X|Y) < CR(Y|X) \) for the whole range of \( \mu \), indicating correctly the direction of the coupling also in this case of two weakly coupled systems.

4. Influence of Observational Noise

In order to study the influence of observational noise on the mean conditional probability of recurrence, we add 25% and 60% observational uniformly distributed white noise to the systems \( X \) and \( Y \) of Eqs. 5,6. Then, we compute \( CR(X|Y) \) and \( CR(Y|X) \) for each value of the coupling strength \( \mu \) (see Figs. 5,6).

With 25% observational noise, we can detect very well the direction of the coupling (Fig. 5). We see that before the onset to synchronization, \( CR(X|Y) < CR(Y|X) \), as expected. With 60% observational noise, the detection of the direction of the coupling is more difficult for very small
Figure 5: \( CR(X|Y) \) (red) and \( CR(Y|X) \) (green) for two identical unidirectionally coupled Henon maps (Eqs. 5,6) contaminated by 25% of observational noise in dependence on the coupling strength \( \mu \).

Figure 6: \( CR(X|Y) \) (red) and \( CR(Y|X) \) (green) for two identical unidirectionally coupled Henon maps (Eqs. 5,6) contaminated by 60% of observational noise in dependence on the coupling strength \( \mu \).

values of the coupling strength \( \mu \) but for \( \mu > 0.1 \) the directionality is correctly detected. Hence, we conclude that the MCR seems to be a rather robust measure for the detection and quantification of the direction of the coupling, also in the presence of high levels of observational noise.

5. Conclusions

In this paper we have introduced a new method to detect and quantify the direction of the coupling in complex systems. The proposed technique is based on the mean conditional probability of recurrence (MCR) and it is applicable to weakly and strongly coupled systems. We have illustrated the applicability of MCR by means of three different numerical examples. Furthermore, we have analyzed the influence of observational noise on MCR and found that it is rather robust. A systematic comparison of MCR with other known techniques for the detection of the directionality is planned for the future, as well as a study of the influence of the threshold \( \varepsilon \) and the length of the time series used for the analysis. Also the problem of passive experiments has to be addressed, i.e. the case in which the time series have been measured for just one value of the coupling strength.

Acknowledgments

The authors would like to thank the "Helmholtz Center for Mind and Brain Dynamics" and the DFG project SPP 1114 for support.

References


Direct or indirect? Graphical models for phase synchronization

Björn Schelter†, Matthias Winterhalder†, Jürgen Kurths‡, Rainer Dahlhaus*, and Jens Timmer‡

† FDM, Freiburg Center for Data Analysis and Modeling, University of Freiburg, Eckerstr. 1, 79104 Freiburg, Germany
‡ Department of Physics, University of Potsdam, Am Neuen Palais, Gebude 19, 14415 Potsdam, Germany
* Institute of Applied Mathematics, University of Heidelberg, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany
Email: schelter@fdm.uni-freiburg.de, winterhm@fdm.uni-freiburg.de, jkurths@agnld.uni-potsdam.de, dahlhaus@statlab.uni-heidelberg.de, jeti@fdm.uni-freiburg.de

Abstract—In many investigations of dynamical systems more than one single process is observed and analyzed. Of particular interest is the detection of interrelationships between processes in such multivariate systems. When more than two processes are analyzed, one has to face the problem that complex interaction structures between the processes may arise. It is not necessary that two processes in a multivariate system have to interact directly. Bivariate analysis is often not sufficient to reveal the correct interaction structure, i.e. distinguishing between direct and indirect interactions. Graphical models applying partial coherence have been introduced to discriminate between spurious and non-spurious interactions in multivariate linear systems. For weakly coupled self-sustained chaotic oscillators it has been observed that these oscillators can synchronize their phases, which can be detected by phase synchronization analysis. A methodology is desired similar to graphical models applying partial coherence to linear systems that is able to distinguish between direct and indirect coupling in non-linear, phase synchronizing systems. To this aim we present the concept of partial phase synchronization. We demonstrate its ability to discriminate direct from indirect interactions in a multivariate system of weakly coupled self-sustained chaotic oscillators.

1. Introduction

In various fields of research analysis techniques have been suggested to analyze interactions in dynamic systems [1, 2, 3, 4, 5, 6, 7]. When more than two processes interact with one another complex interaction structures between the processes may arise, which have to be faced when analyzing such systems. For instance two processes in a multivariate system do not have to interact directly. Therefore, bivariate analysis that neglects contribution of other processes is often not sufficient to reveal the correct interaction structure, i.e. distinguishing direct and indirect interactions. In order to avoid false positive conclusions about the interdependence structure of the investigated multivariate system, the applied analysis technique should distinguish direct and indirect interactions which is impossible for bivariate techniques alone.

Partial spectral coherence was introduced [8] and applied [9] to discriminate direct and indirect connections in linear systems. Graphical models applying partial coherence have been introduced to reveal the interdependence structure in multivariate systems consisting of linear stochastic processes [10]. If a significant bivariate coherence is detected between two processes, which becomes non-significant utilizing partial coherence, the corresponding connection is unmasked as an indirect one. As a result of this, false positive conclusions are prevented in linear systems. An intuitively interpretable graph represents the investigated processes as vertices and direct interrelations are represented by edges between the corresponding vertices.

Here, we carry over the concept of graphical models and partialization analysis to non-linear synchronizing systems. To this aim, we first present the concept of graphical models applying partial coherence and then demonstrate how this methodology can be carried over to phase synchronization analysis. This leads to partial phase synchronization analysis.

2. Graphical models applying partial coherence

Considering an $N$-dimensional dynamic process $X_1,\ldots,X_N$, the partial cross-spectra $S_{X_jX_k|X_l}$ between $X_k$ and $X_l$ and the auto-spectra $S_{X_jX_j}$ of $X_j$ conditioning on all remaining processes $X_Z \{X_Z|Z = 1,\ldots,N, Z \neq k, l\}$ are defined by

$$S_{X_jX_k|X_l}(\omega) = S_{X_jX_k}(\omega) - S_{X_jX_l}(\omega)S^{-1}_{X_lX_k}(\omega)S_{X_kX_l}(\omega)$$

and by

$$S_{X_jX_k}(\omega) = S_{X_jX_k}(\omega) - S_{X_jX_l}(\omega)S^{-1}_{X_lX_k}(\omega)S_{X_kX_l}(\omega).$$

$S_{X_jX_l}(\omega), S_{X_jX_k}(\omega),$ and $S_{X_jX_k}(\omega)$ denote the multivariate auto- and cross-spectra, which can be estimated e.g. by smoothing the corresponding periodograms [11].

Inverting and renormalization of the spectral matrix

$$S(\omega) = \begin{bmatrix} S_{X_1X_1}(\omega) & S_{X_1X_2}(\omega) & \cdots & S_{X_1X_N}(\omega) \\ S_{X_2X_1}(\omega) & S_{X_2X_2}(\omega) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ S_{X_NX_1}(\omega) & \cdots & \cdots & S_{X_NX_N}(\omega) \end{bmatrix}$$

(3)
is a numerically efficient method to estimate the partial auto- and cross-spectra \[10\]. The information about the linear interrelation between the processes \(X_t\) and \(X_t\) conditioned on the remaining examined processes \(X_Z\) is contained in the partial coherence

\[
P_{\text{Coh}}(\omega) = \frac{|S_{X_t|X_1,X_2}(\omega)|}{\sqrt{S_{X_t|X_1}(\omega)S_{X_t|X_2}(\omega)}}. \tag{4}\]

A graphical model reflecting the partial correlation structure consists of a set of vertices \(V = 1, \ldots, N\) and a set of edges \(E\). An edge between \(k\) and \(l\) is present if \(P_{\text{Coh}}(\omega)\) is non-zero \[10\]. We mention that there is a similar concept for ordinary random variables, e.g. covariance selection models \[12, 13\].

As an example, the following vector autoregressive process is considered

\[
x_1(t) = 1.1x_1(t-1) - 0.6x_1(t-2) + 0.5x_2(t-3) + \xi_1(t) \tag{5}
\]

\[
x_2(t) = 0.6x_2(t-1) - 0.4x_1(t-5) + \xi_2(t) \tag{6}
\]

\[
x_3(t) = 0.5x_3(t-3) + 0.4x_2(t-4) + \xi_3(t) \tag{7}
\]

\[
x_4(t) = 1.2x_4(t-1) - 0.7x_4(t-2) - 0.5x_1(t-2) + \xi_4(t) \tag{8}
\]

The analysis of this model system using graphical models applying partial coherence is presented in Fig. 1. In (a) on the diagonal the spectra of the processes are display, while above the diagonal the bivariate coherence is shown and below the diagonal the partial coherence. The resulting graph of this partial coherence analysis is summarized in Fig. 1 (b). The correct interaction structure is revealed by graphical models applying partial coherence. We want to emphasize that ordinary coherence suggests that all processes interact with one another directly.

3. Graphical models applying partial phase synchronization

For nonlinear synchronizing system to investigate phase synchronization

\[
|P_{\text{Coh}}(\omega)| = \frac{1}{T} \sum_{t=1}^{T} e^{j \Phi_{k,l}^m(t)} \tag{9}
\]

has been introduced. Here, we concentrate on \(n = m = 1\). The synchronization index \(P_{\text{Coh}}(\omega)\) quantifies the sharpness of peaks in the histograms of the phase difference \(\Phi_{k,l}^m = m\Phi_k(t) - m\Phi_l(t)\) for appropriate integers \(n\) and \(m\). It is normalized between zero and one. A value close to one is obtained for an almost constant phase difference \(|m\Phi_k(t) - m\Phi_l(t)| = |\Phi_{k,l}^m(t)| < \text{const}, n, m \in \mathbb{Z}\). This has also been observed between two non-identical oscillators \[14\]. For the extension of phase synchronizaion to non-linear stochastic oscillators, the distribution of \(\Psi_{k,l}^m(t) = \Phi_{k,l}^m(t) \mod 2\pi\) is investigated \[4\].

The auto- and cross-spectra enter the spectral matrix

\[
\begin{pmatrix}
R_{1,1} & \cdots & R_{1,N} \\
R_{2,1} & \cdots & R_{2,N} \\
\vdots & \ddots & \vdots \\
R_{N,1} & \cdots & R_{N,N}
\end{pmatrix}
\]

with entries \(R_{i,j} := \Psi_{i,j}^0\) (Eq. 9), which are the pairwise synchronization indices. In the following we refer to this matrix as synchronization matrix. The inverse \(PR = R^{-1}\) of the synchronization matrix \(R\) immediately leads to the definition of the \(1:1\) partial phase synchronization index

\[
R_{k,l,Z} = \frac{|PR_{k,l}|}{\sqrt{PR_{k,k} PR_{l,l}}} \tag{10}
\]

between \(X_k\) and \(X_l\) conditioned on the remaining processes \(X_Z\) (where \(Z \neq k, l\)). For further information and proof, please refer to Ref. [15]. Now it holds that if the bivariate phase synchronization index \(R_{k,l}\) is considerably different from zero, while the corresponding multivariate partial phase synchronization index \(R_{k,l,Z} \approx 0\), there is strong evidence for an indirect coupling between the processes \(X_k\) and \(X_l\). Graphical models applying partial phase synchronization analysis are defined by:

An edge \(E\) between the oscillators \(k\) and \(l\) in a partial phase synchronization graph is missing, if and only if \(R_{k,l,Z}\) is small compared to \(R_{k,l}\).

Three coupled stochastic Roessler oscillators

\[
\begin{align*}
\dot{\xi}_{1,i} &= \frac{X_i}{Y_i} \\
\dot{\xi}_{2,i} &= -\omega_j Y_j - Z_j + \frac{\sum_{i \neq j} \epsilon_{i,j} (X_i - X_j)}{\sqrt{\omega_j X_j + a Y_j}} + \sigma_j \eta_j \\
&= \frac{-\omega_j Y_j - Z_j + b + (X_j - c) Z_j}{\sqrt{\omega_j X_j + a Y_j}} \\
&= i, j = 1, 2, 3
\end{align*}
\]

are a genuine example of a system consisting of weakly coupled self-sustained stochastic oscillators. The parameters are set to \(a = 0.15, b = 0.2, c = 10, \sigma_j = 1.5, \omega_1 = 1.03, \omega_2 = 1.01, \) and \(\omega_3 = 0.99\) yielding a chaotic behavior in the deterministic case. For the noise term \(\sigma_j \eta_j\) a standard deviation of \(\sigma_j = 1.5\) is chosen and \(\eta_j\) is standard Gaussian distributed. Both the bidirectional coupling \(\epsilon_{1,3} = \epsilon_{3,1}\) between oscillator \(\xi_1\) and oscillator \(\xi_3\) and the bidirectional coupling \(\epsilon_{1,2} = \epsilon_{2,1}\) between oscillator \(\xi_1\) and oscillator \(\xi_2\) are varied between 0 and 0.3. Both synchronization phenomena, phase and lag synchronization, are
The oscillators $\xi_2$ and $\xi_3$ are not directly coupled since $\varepsilon_{2, 3} = \varepsilon_{3, 2} = 0$.

The bivariate synchronization index $R_{1, 2}$ as well as $R_{1, 3}$ increases when the corresponding coupling strength is increased, indicating phase synchronization (Fig. 2 (a) – above the diagonal). Once a sufficient amount of coupling exists between oscillators $\xi_1$ and $\xi_2$ as well as between $\xi_1$ and $\xi_3$, a non vanishing bivariate synchronization index $R_{2, 3}$ between the not directly coupled oscillators $\xi_2$ and $\xi_3$ is observed (Fig. 2 (a) – above the diagonal). The common influence from oscillator $\xi_1$ onto $\xi_2$ and $\xi_3$ causes this high but spurious synchronization index. The bivariate synchronization analysis suggests a complete linking between all oscillators, containing the additional but spurious edge between oscillator $\xi_2$ and $\xi_3$.

In Fig. 2(a) (below the diagonal) the results of partial phase synchronization analysis are shown. While $R_{1, 23}$ as well as $R_{1, 32}$ are essentially unchanged compared to the bivariate synchronization indices, $R_{2, 31}$ stays almost always below 0.1 and is therefore considerably smaller than $R_{2, 3}$ in the area of spurious synchronization. This strongly indicates the absence of a direct coupling between oscillators $\xi_2$ and $\xi_3$. This results in the graph presented in Fig. 2 (b), representing the correct coupling scheme.

4. Discussion

In summary, we introduced graphical models applying partial phase synchronization to phase signals of multivariate synchronizing oscillators. This multivariate extension is essential when analyzing multivariate systems to avoid false positive conclusions about the underlying coupling scheme. To this aim, we suggested an index quantifying synchronization by applying a partialization analysis to the analytic signal. Inverting the synchronization matrix yields the concept of partial phase synchronization, that has been shown to differentiate direct and indirect coupling in a multivariate system of non-linear synchronizing oscillators.

References


Figure 2: (a) Phase synchronization and partial phase synchronization index. Coupling strengths between oscillators $\xi_1$ and $\xi_2$ and between oscillators $\xi_1$ and $\xi_3$ are varied between 0 and 0.3, for an absent coupling between $\xi_2$ and $\xi_3$. Values of the bivariate phase synchronization index (above the diagonal) and partial phase synchronization index (below the diagonal) are shown. When comparing the bivariate phase synchronization index $R_{2,3}$ with the partial phase synchronization index $R_{2,3|1}$ it becomes clear that the interaction between oscillator $\xi_2$ and $\xi_3$ is mediated by $\xi_1$ since $R_{2,3} \gg R_{2,3|1}$. (b) Graph for the simulated coupling scheme in the Rössler system. The direct coupling between oscillators $\xi_2$ and $\xi_3$ is absent. The additional but spurious edge between oscillator $\xi_2$ and $\xi_3$ correctly revealed by partial phase synchronization analysis is denoted by the dotted line.


Abstract—In this paper we present an approach to generate (multivariate) twin surrogate (TS) trajectories based on recurrence properties. In contrast to other approaches, such as the linear-like surrogates, this technique generates surrogates which correspond to an independent copy of the underlying system, i.e. they induce a trajectory of the underlying system starting at different initial conditions. We show that these surrogates are well suited to test for complex synchronization. We then apply the TS to study binocular fixational movements and find strong indications that the fixational movements of the left and right eye are phase synchronized. This result indicates that there might be one center only in the brain that produces the fixational movements in both eyes or a close link between two centers.

Twin Surrogates

The concepts of complex synchronization and especially phase synchronization (PS) have been intensively studied in recent years [1]. Indications of PS have been found in many laboratory and natural systems, e.g. population dynamics, neurological, cardio-respiratory systems and optics, or during epilepsy [2]. The corresponding studies are usually based on the computation of a measure which quantifies dependencies of the instantaneous phases of the time series. However, even though these measures may be normalized between 0 and 1, experimental time series often yield values which are neither close to 0 nor to 1 and hence are difficult to interpret. This problem can be overcome if the coupling strength between the two systems can be varied systematically and a rather large change in the measure can be observed (“active experiment”) [1]. PS in natural systems, e.g. during epilepsy or between the heart beats of a mother with the ones of her fetus [3], frequently evades such an experimental manipulation (“passive experiment”). In some cases, this problem has been tackled by interchanging the pairs of oscillators [3], e.g. the heart beats of other pregnant women were used as “natural surrogates”. These surrogates are independent and hence not in PS with the original system. These surrogates mimic the dynamical behavior of the system, i.e. not only the linear properties but also the nonlinear ones are preserved. They correspond to an independent copy of the total system, and hence, they are not in PS with the original one. In the case of stochastic systems they simply correspond to another realization of the process. In the case of deterministic systems, they correspond to a trajectory of the underlying system starting at different (random) initial conditions. Due to the random element in both situations, we will refer to the surrogates as alternative realizations of the respective system. The main idea is to find points of the measured trajectory which are not only neighbors, but also share the same neighborhood in phase space (twins), i.e. all other points are either neighbors or of neither of them. Once the twins of the trajectory have been localized, new surrogate trajectories are generated by substituting randomly the next step in the trajectory by either its own future or the one of its twin. The algorithm to generate the surrogates is based on randomization of the Fourier phases. They mimic the individual spectra of the two components of the original bivariate series as well as their cross-spectrum, i.e. their linear properties, but not the higher order moments. In this case, the corresponding null-hypothesis is that the putative synchronization in the underlying system can be explained by a bivariate linear stochastic process. The specificity of this test is not always satisfactory, because the concept of PS assumes the mutual adaption of self-sustained oscillators, i.e. nonlinear deterministic systems.

In this paper we present a new concept for the generation of surrogates, which is based on the recurrences of a system. These surrogates mimic the dynamical behavior of the system, i.e. not only the linear properties but also the nonlinear ones are preserved. They correspond to an independent copy of the total system, and hence, they are not in PS with the original one. In the case of stochastic systems they simply correspond to another realization of the process. In the case of deterministic systems, they correspond to a trajectory of the underlying system starting at different (random) initial conditions. Due to the random element in both situations, we will refer to the surrogates as alternative realizations of the respective system. The main idea is to find points of the measured trajectory which are not only neighbors, but also share the same neighborhood in phase space (twins), i.e. all other points are either neighbors or of neither of them. Once the twins of the trajectory have been localized, new surrogate trajectories are generated by substituting randomly the next step in the trajectory by either its own future or the one of its twin. The algorithm to generate the surrogates is based on the recurrence matrix

\[ R_{ij} = \Theta(\delta - ||\vec{x}_i - \vec{x}_j||), \]  

where \( \Theta(\cdot) \) denotes the Heaviside function, \( || \cdot || \) the Euclidean norm and \( \delta \) is a predefined threshold. Coding the “1’s” in the matrix as black dots and the “0’s” as white ones, we obtain the recurrence plot (RP) of the trajectory.
It is important to note that if the RP is computed from a univariate time series, it contains all topological information about the underlying attractor, which therefore can be reconstructed from it [6].

Hence, a first idea for the generation of surrogates is to change the structures in an RP consistently with the ones produced by the underlying dynamical system. In this way one could reconstruct a new realization of the trajectory from the modified \( R_{ij} \). However, the structures in RPs are not fully understood and one cannot arbitrarily interchange columns in an RP, because such a modification changes the distribution of diagonal lines and hence the entropy and predictability of the system [6].

Therefore, we propose a modified approach. In general, in an RP there are identical columns, i.e. \( \|x_i - x_j\|_\infty < \delta \), but which also share the same neighborhood. Reconstructing the attractor from an RP, the respective neighborhoods of these points cannot help to distinguish them, i.e. from this point of view they are identical. This is why we will call them **twins**.

The number of twins depends on \( \delta \) and is typically of the order of 10-50% of the length of the time series. Twins are special points of the time series as they are indistinguishable considering their neighborhoods but in general different and hence, have different pasts and -more important-different futures. The key idea of how to introduce the randomness needed for the generation of surrogates of a deterministic system is that one can jump randomly to one of the possible futures of the existing twins.

A surrogate trajectory \( \tilde{x}(i) \) of \( x(i) \) with \( i = 1, \ldots, N \) is then generated in the following way:

(i) Identify all pairs of twins. (ii) Choose an arbitrary starting point, say \( \tilde{x}(1) = \tilde{x}(k) \). (iii) If \( \tilde{x}(k) \) has no twin, the next point of the surrogate trajectory is \( \tilde{x}(2) = \tilde{x}(k+1) \). (iv) If \( \tilde{x}(k) \) has a twin, say \( \tilde{x}(m) \), then one can go to either \( \tilde{x}(k+1) \) or to \( \tilde{x}(m+1) \), i.e. \( \tilde{x}(2) = \tilde{x}(k+1) \) or \( \tilde{x}(2) = \tilde{x}(m+1) \) with equal probability 1. Steps (iii) and (iv) are then iterated until the surrogate time series has the same length as the original one.

This algorithm creates twin surrogates (TS) which are shadows of a (typical) trajectory of the system [7]. In the limit of an infinite original trajectory, its surrogates are characterized by the same dynamical invariants and the same attractor. However, if the measure of the attractor can be estimated from the observed finite trajectory reasonably well, its surrogates share the same statistics. Also their power spectra and correlation functions are consistent with the ones of the original system. TS do not only seem to give reasonable results for deterministic systems; the TS of e.g. an ARMA process also show the typical behavior of a linear Gaussian process if a suitable embedding is used. Even the parameters of the process can be estimated correctly from the surrogates.

Next, we use the TS to test for PS. The idea behind this approach is similar to the one by means of "natural surrogates" in the mother-fetus heartbeat synchronization [3]. Suppose that we have two coupled self-sustained oscillators \( \tilde{x}_1(t) \) and \( \tilde{x}_2(t) \). Then, we generate \( M \) TS of the total system, i.e. \( \tilde{x}_1^{(i)}(t) \) and \( \tilde{x}_2^{(i)}(t) \), with \( i = 1, \ldots, M \). These surrogates are independent copies of the total system, i.e. trajectories of the whole system beginning at different initial conditions.

Note, that the coupling between \( \tilde{x}_1(t) \) and \( \tilde{x}_2(t) \) is also mimicked by the surrogates. Next, we compute the differences between the phases of the original system \( \Delta \Phi(t) = \Phi_1(t) - \Phi_2(t) \) applying e.g. the analytical signal approach [1] and compare them with \( \Delta \Phi^s(t) = \Phi_1(t) - \Phi_2^s(t) \) (one can also consider \( \Phi_1^s(t) - \Phi_2(t) \)). Then, if \( \Delta \Phi(t) \) does not differ significantly from \( \Delta \Phi^s(t) \) with respect to some index for PS, the null hypothesis cannot be rejected and hence, we do not have enough evidence to state PS.

As a test case, we consider two non-identical, mutually coupled Rössler oscillators [1] with a frequency mismatch of \( \nu = 0.015 \). In this "active experiment", we vary the coupling strength \( \epsilon \) from 0 to 0.08 and compute a PS index for the original trajectory for each value of \( \epsilon \). Next, we generate 200 TS and compute the PS index between the measured first oscillator and the surrogates of the second one. As PS index we use the mean resultant length \( R \) of complex phase vectors [8]

\[
R = \left| \frac{1}{N} \sum_{i=1}^{N} \exp(i\Delta \Phi(t)) \right|.
\]

It takes on values in the interval from 0 (non PS) to 1 (perfect PS)[8]. Let \( R^o \) denote the PS index between the observed first oscillator and the surrogate \( i \) of the second one. To reject the null hypothesis at a significance value \( \alpha \), \( R \) must be larger than \((1 - \alpha) \cdot 100 \) percent of all \( R^o \). Fig. 1 a shows the results for \( R \) of the original system (bold line) and the 1% (solid) significance level. Fig. 1 b displays the difference between \( R \) of the original system and the 1%, 2% and 5% significance level. For \( \epsilon < 0.025 \) of the original system, one expects the curves to cross (the difference becomes positive). This is in agreement with the criterion for PS via Lyapunov exponents \( \lambda_i \) [1], i.e. \( \lambda_4 \) becomes negative at \( \epsilon \approx 0.028 \) (Fig. 1 b), which approximately coincides with the intersection of the curve of \( R \) for the original system and the significance level (zero-crossing of the curves in Fig. 1 b). Therefore, we recognize successfully the PS region by means of the TS.

Note, that also the significance limit increases when the transition to PS occurs (Fig. 1 a). As the TS mimic both the linear and nonlinear characteristics of the system, the surrogates of the second oscillator have in the PS region the

---

1 If triplets occur one proceeds analogously.

2 The equations are \( \dot{x}_{1,2} = -(1 \pm \nu)y_{1,2} - z_{1,2} + \epsilon(x_{1,2} - x_{1,2}), \dot{y}_{1,2} = (1 \pm \nu)x_1 + 0.15y_{1,2}, \dot{z}_{1,2} = 0.2 + z_{1,2} + z_{1,2}(x_{1,2} - 10) \)
same mean frequency as the first original oscillator. Hence \( R^0 \) is rather high. However, \( \Phi_1(t) \) and \( \Phi_s^i(t) \) do not adapt to each other, as they are independent. Hence, the value of \( R \) for the original system is significantly higher than the \( R^0 \). We state in conclusion that even though the obtained value for a normalized PS index is higher than 0.97 (right side of Fig. 1a), this does not offer conclusive evidence for PS. Hence, the knowledge of the PS index alone does not provide sufficient evidence for PS. Next, we perform an analysis of the power of the test for

![Figure 1](image1.png)

Figure 1: a) \( R \) of the original data (bold) and significance level of 1% (solid). b) Difference between \( R \) of the original data and significance level of 1% (solid), 2% (dashed) and 5% (dashed-dotted). The zero line is plotted (dotted) to guide the eye. c) Four largest Lyapunov exponents for the 6-dimensional system considered. \( \lambda_4 \) is highlighted and the arrow indicates the transition to PS.

\[ \varepsilon = 0 \text{ and } \nu = 0. \] For 100 random initial conditions of the Rössler system and a significance level of \( \alpha = 1\% \), the null hypothesis was erroneously rejected only in 1 out of the 100 cases. This is a rather auspicious result, as due to the identical frequencies, it is extremely difficult to detect PS in this case. In the case of \( \varepsilon = 0.02 \) (e.g. no PS) and \( \nu = 0.015 \), there were no erroneous rejections of the null hypothesis. Finally, for PS (\( \varepsilon = 0.045 \) and \( \nu = 0.015 \)), in all 100 test runs the null hypothesis was correctly rejected. These results indicate that the power of the test is good. Next we apply our algorithm to test fixational movements of the two eyes for PS. During fixation of a stationary target our eyes perform small involuntary and allegedly erratic movements to counteract retinal adaptation. If these eye movements are experimentally suppressed, retinal adaptation to the constant input induces very rapid perceptual fading [9]. Moreover, statistical correlations show a timescale separation from persistence to antipersistence [10]. Persistence on the short timescale counteracts retinal fading, whereas antipersistence on the long timescale contributes to stability of ocular disparity. According to current textbook knowledge, the fixational movements of the left and right eye are correlated very poorly at best [11]. Therefore, it is highly desirable to examine these processes from a perspective of PS. We analyze the data of two subjects. Each performed three trials, in which they fixated a small stimulus (black square on a white background, \( 3 \times 8 \) pixels on a computer display) with a spatial extent of 0.12°, or 7.2 arc min. Eye movements were recorded using an EyeLink-II system (SR Research, Osgoode, Ontario, Canada) with a sampling rate of 500 Hz and an instrument spatial resolution less than 0.005°. Fig. 2 shows a segment of the horizontal (a) and vertical (b) component of the eye movements for one person. The data were first high-pass filtered applying a difference filter \( \tilde{x}(t) = x(t) - x(t - \tau) \) with \( \tau = 40 \text{ ms} \) in order to eliminate the slow drift of the data. After this filtering, we find an oscillatory trajectory (Fig. 3), which has maximum spectral power in the frequency range between 6 and 8 Hz. However, the trajectories of the eyes are rather noisy and non-phase coherent. Therefore, it is cumbersome to estimate the phase of these data. Hence we apply another measure of PS which is based on the probability of recurrence of a trajectory in phase space \( P(\tau) = 1/N \sum_{i=1}^{N} R_{i+\tau} \), where \( R_{i+\tau} \) is the recurrence matrix (Eq. 1). The correlation between the probabilities of recurrence of two interacting oscillators

\[ CPR = \frac{P_{1}(\tau)}{P_{2}(\tau)} > \frac{\sigma_{1} \sigma_{2}}{\sigma_{1} \sigma_{2}}, \]

(3)

where \( \bar{P}_{1,2} \) means that the mean value has been subtracted and \( \sigma_{1} \) and \( \sigma_{2} \) are the standard deviations of \( P_{1}(\tau) \) respectively \( P_{2}(\tau) \), has been proposed to detect PS in non-phase coherent and noisy oscillators, where the phase cannot be estimated directly [12]. Now, we compute 200 surrogates (Fig. 3b) of the left eye’s trajectory and compute the recurrence based synchronization index \( CPR \) between them and the measured right eye’s trajectory. In Fig. 3 c) the results of the test of one trial are shown. In all cases, the PS index of the original data is significantly different from the ones of the surrogates, which strongly indicates that the concept of PS can be succesfully applied to study the interaction between the trajectories of the left and right eye during fixation. This result also suggests that the physiological mechanism in the brain that produces the fixational eye movements controls both eyes simultaneously, i.e., there might be only one center in the brain that produces the fixational movements in both eyes or a close link between two centers. Our finding of PS between left and right eyes is in good agreement with current knowledge of
the physiology of the oculomotor circuitry. In a single-cell study, 66% of abducens motor neurons fired in relation to the movements of either eye, while premotor neurons in the brainstem encode monocular movements [13]. Thus, motor neurons—as the final common pathway of neural control of eye movements—are candidates for the synchronization of left and right fixational movements. Furthermore, we are interested in whether the fixational movements in the horizontal and vertical direction of one eye are synchronized. Horizontal and vertical saccadic eye movements are controlled in two spatially distinct brainstem nuclei [14]. Therefore, we can expect that, on the level of fixational eye movements, horizontal and vertical components are independent. Applying the method of TS to this case, we find that the synchronization index CPR between the x- and y-component is not significantly different from the one computed for the surrogates. Hence, we do not have sufficient evidence to claim synchronization between the horizontal and vertical components of eye movements.

In conclusion, we propose a new method for generating surrogates based on the concept of recurrence. These TS correspond to an independent copy of the underlying system, i.e. to a trajectory of the system starting at different initial conditions. We have shown that the TS can be used to test for PS in the well studied system of two mutually coupled Rössler oscillators, and, by means of the TS, we have detected PS in dependence of the coupling strength at several significance levels. Furthermore, we have tested for PS in experiments of binocular fixational movements and found that the left and right eye are in PS, in agreement with physiological results about the functional role of motor neurons in the final common pathway for the control of eye movements. Contrary to popular belief, fixational eyes movements are a necessary condition for vision. Thus, an understanding of their dynamics is fundamental for perception and the associated control of spatial attention [15].

First results indicate that the concept of TS can be also used for testing other complex kinds of synchronization, especially generalized synchronization.

Acknowledgments

This work has been supported by the DFG priority program 1114, DFG-KL 955, and the "Internationales Promotionskolleg - Helmholtz Center for the Study of Mind and Brain Dynamics" at the University of Potsdam.

References

Assessing Dependences within Multivariate Time Series Partializing the Knowledge of Thirds

Cristian Carmeli† and Oscar De Feo‡

†Laboratory of Nonlinear Systems, Swiss Federal Institute of Technology Lausanne; EPFL-IC-LANOS, Station 14, CH-1015 Lausanne, Switzerland
‡Department of Microelectronic Engineering, University College Cork; Butler Building, North Mall, Cork, Ireland
E-mail: cristian.carmeli@epfl.ch; oscar.defeo@ucc.ie

Abstract—A method to estimate from multivariate measurements the dependences within a network of coupled dynamical systems is proposed. The method is non-parametric and resorts to a statistics of the eigen-spectrums of the time series partial correlation matrices. The method is successfully validated on numerically generated data, demonstrating its capability to distinguish between direct and indirect dependences.

1. Introduction

In modern experimental setups, with the growing availability of multiple parallel measurements, multivariate time series analysis has been raising as a key topic. In particular, inferring from measured data the functional interactions (dependences) between coupled dynamical systems has become a crucial step to unravel the principles governing the observable collective behaviors of concrete networks [1]. For instance, in population biology the interest is focused on interactions between different populations in a given territory [2], while in neuroscience a key question is how single neurons process a certain stimulus within functionally coherent neuronal assemblies [3].

Traditionally, this step consists in estimating the strength of dependences between measured signals (time series) with estimators such as linear cross-correlation, mutual information, measures of mutual predictability and phase dynamics [4, 5]. Unfortunately, these estimators are usually not viable in a multivariate context, either because of their computational complexity or because their inability to marginalize third knowledge. Indeed, when dealing with multivariate data, the computational costs might explode and direct and indirect dependences must be discerned [6].

An effective dependences estimator for multivariate stochastic processes is already available [6], while for deterministic processes contributions have started to appear only very recently (e.g. [7]).

Here, we present an approach to estimate the strength of dependences between multivariate time series that, theoretically, can be applied to both deterministic and stochastic processes. The method is non-parametric and represents an extension of a recently proposed multivariate dependences estimator [8].

The method is described in Sec. 2; in Sec. 3 its ability to correctly estimate dependences within networks of three dynamical systems is assessed on numerically generated data; finally, conclusions are given in Sec. 4.

2. Method

We first briefly recall the recently introduced dependence estimator, called $S$ [8], then we explain how to correctly apply it on measurements from deterministic dynamical systems, and, finally, we will extend it to estimate dependences between two time series conditionally upon third ones.

2.1. $S$ Estimator

This estimator has been recently introduced to quantify the cooperativeness within a network of dynamical systems out of measured time series [8].

Given $M$ time series, one from each dynamical system under study, we denote them by $\mathbf{Y} = \{y_i\}, i = 1, \ldots, L$, where $y_i \in \mathbb{R}^M$ is the $i$-th sample observation vector and $L$ is the number of available samples. Without loss of generality, we can consider $\mathbf{Y}$ as de-trended to zero mean and normalized to unitary variance.

Let us consider the $M \times M$ estimated correlation matrix of the time-series

$$
\mathbf{C} = \frac{1}{L-1} \sum_{i=1}^{L} \mathbf{y}_i^T \mathbf{y}_i ,
$$

with elements $c_{ii} = 1$ and $c_{ij} = \rho_{ij}, i \neq j$, i.e. the correlation between the $i$-th and $j$-th time series. Called $\lambda'_{i} = \frac{1}{M} \mathbf{C}$ the normalized eigenvalues of $\mathbf{C}$, the entropy-like quantity

$$
H = - \sum_{i=1}^{M} \lambda'_i \log (\lambda'_i)
$$

is a measure inversely proportional to the amount of dependences between the $M$ time series.

Indeed, it can be interpreted as a deviation from mutual orthogonality (lack of correlation) between the $M$ signals.
In the case of $M$ uncorrelated signals, $C = I$, the normalized eigenvalues are all equal $\lambda'_i = \frac{1}{M}$, and $H$ is equal to $\log M$. In the case of perfectly correlated signals, $C$ has one unitary normalized eigenvalue and all the others zero, and $H$ is equal to 0. To have a measure proportional to the amount of dependences, we can simply rearrange Eq. (2) as $S = 1 - \frac{H}{\log M}$, which is 0 for uncorrelated signals, 1 for completely correlated ones, is monotonically increasing with respect to all correlation terms (cf. [8]), i.e. the off-diagonal elements of $C$, and has been shown to scale with coupling strength when considering coupled deterministic non linear dynamical systems [8].

By reconstructing from the measured scalar time series, through embedding, the trajectory of the dynamical phenomena under observation, this estimator can explicitly account for the hypothesis of deterministic dynamical systems behind the measurements. However, in this case a normalizing step is necessary.

Given, for the sake of simplicity, two time series ($M = 2$), for which delay times ($\tau^{(1)}$ and $\tau^{(2)}$) and embedding dimensions ($m^{(1)}$ and $m^{(2)}$) have been estimated [5], we can consider the embedded multivariate trajectory $X = \{x_i\}$, where $x_i \in \mathbb{R}^{m^{(1)} + m^{(2)}}$ and

$$x_i = \begin{bmatrix} y_i^{(1)}(\tau^{(1)}), & \ldots, & y_i^{(1)}(\tau^{(1)} + (m^{(1)} - 1)\tau^{(1)}), & y_i^{(2)}(\tau^{(2)}), & \ldots, & y_i^{(2)}(\tau^{(2)} + (m^{(2)} - 1)\tau^{(2)}) \end{bmatrix}.$$  

The corresponding estimated correlation matrix can be block partitioned to highlight the contribution of the two systems, i.e.

$$C = \begin{bmatrix} C^{(1,1)} & C^{(1,2)} \\ C^{(2,1)} & C^{(2,2)} \end{bmatrix},$$  

where the $m^{(1)} \times m^{(1)}$ matrices $C^{(i,j)}$, $i = 1, 2$, collect the intra-system correlation terms, i.e. the correlation between state-variables of the same system, while the $m^{(1)} \times m^{(2)}$ matrix $C^{(1,2)}$ collects the inter-system correlation terms, which are the dependences in our interest.

To correctly estimate the interdependence between the two systems, independently of the intra-dependences, we proceed through a suitable linear transformation of the reconstructed state space $X$ which reduces the $C^{(i,j)}$ to identity matrices; in other words, we intra-orthogonalize the state variables of the two systems. As result of the transformation, the estimated correlation matrix for the transformed trajectory will have nonzero off-diagonal elements only within the inter-dependence block $C^{(1,2)}$.

The transformation is given by

$$Z = \begin{bmatrix} Z^{(1)} \\ Z^{(2)} \end{bmatrix} = \begin{bmatrix} X^{(1)} & X^{(2)} \end{bmatrix} \begin{bmatrix} T^{(1)} & 0 \\ 0 & T^{(2)} \end{bmatrix} = XT,$$

with $T^{(1)} = C^{(1,1)^{-\frac{1}{2}}}$ and $T^{(2)} = C^{(2,2)^{-\frac{1}{2}}}$, i.e. the principal square root matrices of $C^{(1,1)^{-1}}$ and $C^{(2,2)^{-1}}$ respectively, where the inverses are guaranteed to exist if an appropriate embedding is performed. Clearly, the estimated correlation matrix for the $Z$ trajectory turns out to be

$$R = \begin{bmatrix} I & R^{(1,2)}(2,1) \\ T^{(2)} & C^{(2,2)}T^{(1)} \end{bmatrix} = \begin{bmatrix} I & R^{(1,2)} \\ R^{(1,2)^T} & I \end{bmatrix},$$

which can then be used, through Eq. (2), to correctly quantify the dependence between the two systems. If the two systems are uncorrelated, $R^{(1,2)} = 0$, $R$ will be diagonal, and $H = \log (m^{(1)} + m^{(2)})$, while if the two systems are “identical”, it can be verified that $R^{(1,2)}$ will have ones on the main diagonal and zeros elsewhere. Under this case the entropy of the normalized eigenvalues will depend on the embedding dimensions ($m^{(1)}$ and $m^{(2)}$). We do not have a closed form formula to compute it; though, it can be easily computed numerically and we denote with $H_{\text{max}}$ its value.

Knowing the entropies, we can finally rearrange Eq. (2) as

$$S = \frac{\log (m^{(1)} + m^{(2)}) - H}{\log (m^{(1)} + m^{(2)}) - H_{\text{max}}}.$$

getting a measure proportional to the amount of dependences and ranging from 0 to 1. Clearly, this procedure can be trivially extended to estimate the whole cooperativeness within $M$ interacting dynamical systems.

### 2.2. Partial $S$ estimator

In the previous section we have derived an estimate for the cooperativeness strength between coupled deterministic dynamical systems. However, this derivation is still not satisfactory because cooperativeness between systems may be merely incidental to the fact that all systems may be commonly correlated to another system. This difficulty can be overcome by considering the partial correlation matrix instead of the correlation one. Indeed, partial correlations are a well established statistical tool to examine correlations between signals conditionally upon thirds signals [9].

To illustrate the procedure, let us consider three deterministic systems, assuming that we want to estimate the dependence between the first two marginalizing the knowledge of the third one.

After a suitable embedding, the $M \times M$ (where $M = \sum_{i=1}^{3} m^{(i)}$) correlation matrix $C$ of the embedded trajectory is estimated. $C$ can be tri-partitioned similarly as Eq. (3) and, accounting for the partitioning, the $(m^{(1)} + m^{(2)}) \times (m^{(1)} + m^{(2)})$ partial correlation matrix between the first two systems given the third one is written as [9]

$$P = \begin{bmatrix} P^{(1,1)} & P^{(1,2)} \\ P^{(2,1)} & P^{(2,2)} \end{bmatrix} = \begin{bmatrix} C^{(1,1)} & C^{(1,2)} \\ C^{(2,1)} & C^{(2,2)} \end{bmatrix} \begin{bmatrix} C^{(1,3)^{-1}} & C^{(1,3)^{-1}C^{(3,2)}} \\ C^{(2,3)^{-1}} & C^{(2,3)^{-1}C^{(3,2)}} \end{bmatrix}.$$  

From this point, we can proceed as in the previous section. Firstly, through a linear transformation of the state
3. Numerical Validation

The $S$ and $pS$ estimators have been validated on numerically generated data. We considered the two heterogeneous networks of Fig. 1, which are composed of three structurally different dynamical systems; namely, a Rössler (R), Lorenz (L) and Colpitt’s (C) oscillator. These two coupling schemes are exemplary of the difficulties arising when estimating interactions among multivariates. In scheme (a) a non-existent (indirect) connection between L and C may be inferred because of the common source R, also called a confounder. In scheme (b), a non-existent interaction between R and L may be inferred because of their common destination C, or child.

The equations governing the dynamics of the two considered networks are given by

$$
\begin{align*}
\dot{\theta}_1^{(1)} &= T \left[ \theta_1^{(1)} + \alpha \theta_1^{(2)} + \theta_1^{(3)} \right] + \sigma \eta_1^{(1)}, \\
\dot{\theta}_2^{(1)} &= T \left[ \theta_1^{(1)} + \alpha \theta_1^{(2)} + \theta_1^{(3)} \right], \\
\dot{\theta}_1^{(2)} &= T \left[ \theta_1^{(3)} - \theta_1^{(2)} + \eta_1^{(2)} \right], \\
\dot{\theta}_1^{(3)} &= T \left[ \theta_1^{(1)} - \theta_1^{(2)} + \eta_1^{(3)} \right], \\
\dot{\theta}_2^{(2)} &= T \left[ \theta_1^{(2)} - \theta_1^{(2)} + \eta_1^{(2)} \right], \\
\dot{\theta}_2^{(3)} &= T \left[ \theta_1^{(3)} - \theta_1^{(2)} + \eta_1^{(3)} \right],
\end{align*}
$$

where $\theta_1^{(1)}, \theta_1^{(2)}, \theta_1^{(3)} j = 1, 2, 3$ are the state variables of the Rössler, Lorenz and Colpitts oscillators, respectively;

where $\theta_1^{(1)}, \theta_1^{(2)}, \theta_1^{(3)} j = 1, 2, 3$ are the state variables of the Rössler, Lorenz and Colpitts oscillators, respectively; $\sigma$ is the standard deviation of the Gaussian noise.$^1$

$^1$In reality the similarity transformation $P = T^TPT$, with $T = \begin{bmatrix} P_1^{(1,3)}, & P_2^{(1,3)} \\ P_3^{(1,3)} \end{bmatrix}$ can be applied directly to the matrix $P$.

Figure 1: Networks of coupled Rössler (R), Lorenz (L) and Colpitt’s (C) dynamical systems considered for the validation of the method. Cases of triangular dependences: (a) – common source; (b) – common child.

Figure 2: Estimated $C-L$ dependence in the case of common source connection (cf. Fig. 1(a)): dependences of (a) – $pS^{(2,3)}$ and (b) – $S^{(2,3)}$ upon the coupling strengths $K^{(2,1)}$ and $K^{(3,1)}$.

Figure 3: Estimated $R-C$ and $R-L$ dependences in the case of common source connection (cf. Fig. 1(a)): dependences of (a) – $pS^{(1,2)}$ and (b) – $pS^{(1,3)}$ upon the coupling strengths $K^{(2,1)}$ and $K^{(3,1)}$. 

$\begin{align*}
a &= 0.4, \ b &= 0.4, \ c &= 5.7, \ \sigma &= 10, \ \beta &= 8/3, \ \tau &= 28, \\
g &= 10^{0.625}, \ Q &= 10^{0.15}, \ \alpha &= 0.996, \ k &= 0.5 \text{ are standard valued parameters; the time scale } T = 6 \text{ adapts the relative speed differences between the three sub-systems; } \eta_1^{(i)}, \ i, j = 1, 2, 3, \text{ are zero mean uncorrelated Gaussian random noise (set to at a strength of 1% of the energy of the right hand side along the uncoupled attractors); and the } K^{(2,1)} \text{ and } K^{(3,1)} \text{ and } K^{(3,2)} \text{ are the diffusive coupling strengths corresponding to the situations of Fig. 1, which in the simulations have been varied within the intervals } [0, 15] \text{ (the former) and } [0, 5] \text{ (the latter two).} \\
\text{For every considered value of the couplings the network was simulated starting from random initial conditions. The transients were discarded, and time series of length } L = 5000 \text{ were collected by sampling } (\delta T = 0.02) \text{ the coupled variables } \theta_1^{(1)}, \ \theta_1^{(2)} \text{ and } \theta_1^{(3)} \text{ corrupted by zero-mean white Gaussian observational noise leading to } 40 \text{ } dB \text{ SNR. From these measurements, we reconstructed 4-dimensional state spaces by delay embedding the time series with } \tau^{(i)} = 0.18, \ i = 1, 2, 3. \\
\text{The results for the case of common source (Fig. 1(a)) are shown in Figs. 2 and 3. Figure 2 reports the dependence of } pS^{(2,3)} \text{ and } S^{(2,3)} \text{ upon the coupling strengths } K^{(2,1)} \text{ and } K^{(3,1)} \text{ evaluated at 100 evenly spaced points. Correctly, } pS^{(2,3)} \text{ stays close to zero and do not scale with neither of the coupling strengths, whilst the } S^{(2,3)} \text{ does scale, showing that the marginalization upon the mea-}
\end{align*}$
measurements from third system does improve the dependence estimation. Moreover, $pS^{(2,3)}$ is always inferior to $pS^{(1,2)}$ and $pS^{(1,3)}$, which, as shown in Fig. 3, scale correctly with the coupling strengths. Also, we remark that they decrease slowly with the increase of both coupling strengths. This phenomenon can be explained by the fact that the three systems influence each other and, consequently, become more and more similar, jeopardizing the reconstruction.

As a consequence of these results, one could think to consider only partial $S$ for estimating dependences, disposing of $S$. However, this is not the case because of the so-called “marrying-parents” effect, commonly observable in the case, illustrated in Fig. 1(b), of a common child. For this case, as shown in Fig. 4, $pS^{(1,3)}$ and $pS^{(2,3)}$ scale correctly with the coupling strengths $K^{(3,1)}$ and $K^{(3,2)}$. However, as shown in Fig. 5(a), $pS^{(1,2)}$ does scale with the couplings, leading to the incorrect inference of a non-existent dependence between $R$ and $L$. Though, the voidance of this coupling can be easily tested by means of the $S^{(1,2)}$ estimator which, as shown in Fig. 5(b), correctly stays close to zero.

From these two numerical experiments we can conclude that, by combining both $S$ and $pS$ estimators, we can correctly estimate the dependences within a network of coupled dynamical systems.

4. Conclusions

A new method to infer from measured time series the strength of interactions within a network of coupled dynamical systems has been proposed, and its ability to discriminating direct from indirect dependences has been demonstrated on numerical data.

The method proved eligible for the application on experimental data. However, toward a significant and appropriate experimental application, deeper and extensive simulation studies are needed. This is matter of ongoing research and will be presented in a later work.

Acknowledgments

The authors thank R. Costantini for fruitful discussions and acknowledge financial support from the European project BRACCIA (NEST-517133).

References

Investigating the limitations of compartmental disease models with Kalman filters

David M. Walker†, Heung Wing Joseph Lee‡ and Michael Small‡

†Department of Applied Mathematics, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
‡Department of Electronic & Information Engineering, Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Email: enwalker@eie.polyu.edu.hk, majlee@polyu.edu.hk, ensmall@eie.polyu.edu.hk

Abstract—We investigate the limitations of traditional discrete-time epidemiological models of disease transmission using a novel application of Kalman filtering. In particular, the performance of the recently developed KFC algorithm is compared to standard extended Kalman filtering for joint state and parameter estimation of S(E)IR models using time series data from the SARS epidemic. The use of Kalman filters suggests a way in which S(E)IR models can be generalized. The limitations of these generalizations to adequately model the spread of SARS is tested using surrogate data analysis.

1. Introduction

The aim of this paper is twofold. We want to investigate how well a recent variant of Kalman filtering performs during simultaneous state and parameter estimation problems. We also want to test the limitations of standard compartmental epidemiological models, and generalizations thereof, to capture the dynamics of the spread of an actual epidemic.

The SARS outbreak in Hong Kong was tragic for some individuals and generally disruptive for the wider community. The effect of an epidemic on the population can be mitigated by control measures such as quarantine. The appropriateness of control measures can be determined, evaluated and decided upon by officials through simulation studies of transmission dynamic models of a virus. It is therefore important to develop and investigate models of disease transmission.

It has been demonstrated that mathematical epidemiological models in their standard form are not appropriate. Even stochastic generalizations have been shown to be lacking depending on the way the standard models are modified. For example, Small et al. [5] added noise to system (difference) equations of SIR models and showed the behaviour to be inconsistent with the data of the SARS outbreak.

A generalization to compartmental disease models not made was to allow the parameters of the models to vary by day or other time period. It is unclear how such stochasticity could be managed. In this work, however, we apply Kalman filters to gather estimates of the parameters sequentially. The parameters are thus estimated adaptively and we propose to use the varying Kalman filter parameter estimates as a distribution from which to draw sensible (daily) varying values. That is, we use the results of our Kalman filter studies to propose a (stochastic) generalization to (discrete-time) compartmental models. We use a simple surrogate/randomization test to compare whether simulated realizations of these models are consistent with the actual recorded SARS outbreak data in Hong Kong.

2. Compartmental Disease Models

The dynamics of the spread of an infectious disease among a population has been studied using compartmental models by Kermack and McKendric [3]. In such models the total population is divided into compartments reflecting the nature of the individuals with respect to the epidemic.

Two commonly studied compartmental models are the so-called SIR and SEIR models. That is, the susceptible-infected-removed population model, and the susceptible-exposed-infected-removed model. The susceptible compartment represents those individuals in the population susceptible to catching the disease. The infected, or infectious, compartment consists of those individuals that have caught the disease and through contact with individuals in the susceptible compartment can propagate the disease. The individuals in the removed compartment have either recovered from the disease and are no longer infectious, or have fallen to the disease. In either case through immunity or death they are no longer considered susceptible. (There are other models such as SIRS where recovered individuals can lose immunity and become susceptible to the disease once more.) The exposed compartment in the SEIR model represents those individuals that have been infected by the disease but are as yet not infectious. That is, the exposed compartment attempts to capture the latent period of an infectious disease.

The dynamics of such models are typically described by simple ordinary differential equations modelling the transmission rates between the various compartments. See, Murray [4], for a good introduction. In this paper we con-
sider the corresponding difference equations obtained from such differential equations.

2.1. SIR model

A discrete-time SIR model can be represented by the following equations [1]

\[
\begin{bmatrix}
S_{n+1} \\
I_{n+1} \\
R_{n+1}
\end{bmatrix} = \begin{bmatrix}
S_n(1 - \frac{\alpha}{\mu} I_n) \\
I_n(1 - \beta + \frac{\mu}{\gamma} S_n) \\
R_n + \beta I_n
\end{bmatrix}
\]

(1)

The total population is subject to the constraint \( L = S + I + R \). (We take \( L \) to be roughly the population of Hong Kong, i.e., \( L = 7 \) million.)

The reproductive rate \( R = \frac{\beta}{\mu} \) gives important information on the development of a potential epidemic. In particular \( R \leq 1 \) implies no epidemic as the rate of infection is no larger than the rate of recovery so eventually the number of new infections decreases.

2.2. SEIR model

The SEIR model generalizes the SIR model by including a new state (exposed compartment) to account for the latent period of diseases. A version of a discrete-time SEIR model is given by the following equations

\[
\begin{bmatrix}
S_{n+1} \\
E_{n+1} \\
I_{n+1} \\
R_{n+1}
\end{bmatrix} = \begin{bmatrix}
S_n(1 - \frac{\alpha}{\mu} I_n) \\
E_n(1 - \gamma) + \frac{\alpha}{\mu} S_n I_n \\
I_n(1 - \beta) + \gamma E_n \\
R_n + \beta I_n
\end{bmatrix}
\]

(2)

where the total population is again constrained by \( L = S + E + I + R \). The new parameter \( \gamma \) accounts for the rate an infected individual becomes infectious and \( \alpha \) is re-interpreted accordingly.

3. Kalman Filtering

The extended Kalman filter (EKF) sequentially processes time series data to produce a (statistical) estimate for the system state. The system state is assumed to be a Gaussian random variable \( x_n \sim N(\mu_n, \Sigma_n) \) and under suitable assumptions about the noise corruption (independent additive Gaussian in both the system dynamics and measurement function) recursively updates \( \mu_n \) and \( \Sigma_n \) in such a way that they are the minimum mean square error estimates. In the case of linear systems and measurement functions the Kalman filter provides the optimal solution but unfortunately in the nonlinear case the EKF is a sub-optimal, but still very useful, heuristic.

The Kalman filter framework can also be used for joint (or simultaneous) state and parameter estimation, so that system parameters can be adaptively fitted to new data as it is processed. The equations in such a framework are as follows:

\[
\begin{align*}
x_{n+1} &= F[x_n, \mathbf{a}_n] + \xi_n \\
\mathbf{a}_{n+1} &= \mathbf{a}_n + \eta_n \\
y_n &= G[x_n, \mathbf{a}_n] + \nu_n
\end{align*}
\]

(3)

where the bracketed equations on the left represent the system dynamics and the adaptive update of the parameters. The equation on the right represents the observation function producing the observed time series. The Greek symbols represent the Gaussian noise corruption at various levels \( (\xi_n \sim N(0, \mathbf{Q}_n), \eta_n \sim N(0, \mathbf{Q}_p) \) and \( \nu_n \sim N(0, \mathbf{R}_n) \)). The way in which the Kalman filter updates the recursive estimates can be found in many textbooks and papers, e.g., see Jacobs [2].

Since the EKF is not an optimal solution for recursive state estimation in the nonlinear case many variants have been proposed to improve performance in different situations. One promising modification is the recently developed KFC algorithm [6, 7] (Kalman filter with constraints) which proposes including additional dynamical information as constraints to be considered an addendum to the observation model framework of the EKF. For example, in [6, 7] dynamical information such as the location of unstable fixed points were used as extra observation functions to improve parameter fitting of nonlinear models with the Kalman filter. Further applications of the KFC algorithm include using it to detect candidate fixed points from data [8].

The KFC is similar to the EKF and the recursive estimates are updated in the same way. The framework for the KFC for joint state and parameter estimation is given by

\[
\begin{align*}
x_{n+1} &= F[x_n, \mathbf{a}_n] + \xi_n \\
\mathbf{a}_{n+1} &= \mathbf{a}_n + \eta_n \\
y_n &= G[x_n, \mathbf{a}_n] + \nu_n
\end{align*}
\]

(4)

where we now have a noise term associated with the introduction of the constraint function \( C[x_n, \mathbf{a}_n] \), i.e., \( \epsilon_n \sim N(0, \mathbf{R}_c) \).

The disease models above are subject to a total population constraint and so an analysis using the KFC seems more appropriate than just using the EKF. We investigate the difference in results between the two algorithms in the next section by using the published case data of the SARS outbreak in Hong Kong.

There is a third variant where we return to an EKF but rewrite the SIR constraint as \( L = S_n + I_n + R_n \) and substitute into the dynamic equations. (Similarly for the SEIR model.) In this way we can study if it is better to “increase” the nonlinearity of the EKF problem by writing the constraint as part of the dynamics, or increasing the dimension of the observational model as is done in the KFC.

4. Model Fitting and Testing

In this section we attempt to fit the parameters \( (\alpha, \beta, \gamma) \) of the SIR model and \( (\alpha, \beta, \gamma) \) of the SEIR model using joint state and parameter estimation with Kalman filters for the SARS epidemic. The available data for the spread of SARS in Hong Kong is the reported number of new infected individuals per day. This data is therefore a proxy for the observations \( y_t = \frac{a}{\mu} S(t) \) for each disease model. This in-
interpretation of the data is reasonable, particularly so, because a feature of the Hong Kong outbreak was nosocomial transmission. If, however, there were no nosocomial transmission the recorded data would actually be the number of new removals from the population and $\beta I$, since identified new infectives were hospitalized and removed from the general population. We do not consider this possibility here for reasons of space but it is straightforward to incorporate into the Kalman filter frameworks.

The Kalman filter requires initialization of the system state and specification of the (design) parameters $Q$ and $R$. We have set the initial state $z_0 = (L-1, 1, 0)$ and initial parameters $(\alpha_0, \beta_0) = (0, 0)$ for the SIR model. For the SEIR model we use $z_0 = (L-1, 0, 1, 0)$ and $(\alpha_0, \beta_0, \gamma_0) = (0, 0, 0)$. In all examples we use $Q = 10^{-4} \sigma^2_I$, $R = 10^{-4} \sigma^2_C$ and $R_c = 10^{-8} \sigma^2_y$, where $I$ is an identity matrix of appropriate size and $\sigma^2_y$ is the square of the standard deviation of the data set.

4.1. SIR model

The results of applying the EKF algorithm to the reported SARS case data are shown in Figure 1. We see that the observed data can be tracked well using the Kalman filter estimates (of both state and parameters). Closer inspection, however, reveals that the constraint on the population is not even close to being satisfied. As such the values of the estimated parameters even after convergence of the filter (around day 10) should be treated suspiciously.

![Figure 1: The EKF state and parameter estimates of the SIR model together with the fit to the observed data, the population constraint and the adaptively evolving reproductive rate.](image)

We next apply a KFC to the data where the population constraint is explicitly included as an extra observation. We show the results of this algorithm in Figure 2. We note that as was the case with the EKF the state and parameter estimates are able to track the observed data very well (after convergence around day 10). In this case however we see that the population is more closely satisfied and so we have greater confidence in the corresponding parameter estimates. We also see that after (apparent) convergence the reproductive rate is approximately 1 suggesting that the disease will eventually die out as indeed it does.

The ability of the KFC to preserve the population constraint implies that it is a better algorithm than the EKF for joint state and parameter estimation. We also note but do not show that an EKF with the relation $L = S_n + I_n + R_n$ substituted directly into the dynamics does not improve the performance of the filter with similar violation of the population constraint to the EKF being seen.

![Figure 2: The KFC state and parameter estimates for the SIR model. We also show the adaptive fit to the data and how the population constraint is met throughout the simulation. The adaptively changing reproductive rate is also displayed.](image)

4.2. SEIR model

The results for the SEIR model are similar to those seen for the SIR model. However, things are much worse and the increase of complexity - higher dimensional systems and an additional parameter - means the results should be treated with suspicion. For reasons of space we only show the results for the KFC algorithm. As expected they are better than those obtained using an EKF. In fact, the EKF produces states estimates with the population in the removed compartment being negative! We can see from Figure 3 that there are some problems, in particular the reproductive rate which is much higher than 1. This suggests that the values estimated for $\alpha$ is far too high. We therefore strongly suspect that the SEIR is an inappropriate model for the SARS data even when the parameters are allowed to adaptively change.

4.3. Stochastic SIR model

The capability of the Kalman filter to track the observed data by allowing the contact rates to adaptively change sug-
gests a way in which the standard disease models considered can be generalized. Specifically, we allow the rates to change each day with the value being sampled from the estimated distribution given by the Kalman filter estimates. (Examining Figure 2 we select both $\alpha$ and $\beta$ to be uniform distributed over [10, 13].) A daily variation of the contact loses the physical interpretation of the parameter as a contact rate, however, we can think of a variation of the parameter as an approximation to a network description of the population where some individuals have more associates than others and hence effectively have a higher contact rate than others. We test the consistency of 100 realizations of the model with the original data using the auto-correlation at lag 1. We can see from Figure 4 that generalizing the SIR model by allowing for (daily) stochastic variation in parameters still does not result in simulated data consistent with the SARS outbreak.

5. Conclusion

We investigated the inability of discrete-time compartment disease models to fully describe the behaviour of epidemics, in particular the SARS outbreak in Hong Kong during 2003. A key point of failure of these models is the assumption of uniform contact rate. We used Kalman filters to consider adaptive contact rates. A standard EKF was seen to be deficient and in the SIR case this could be corrected using a KFC with the constraint on the population explicitly considered. The Kalman filter estimates appeared to fit the observed data well which suggested a way of generalizing the models: a stochastic generalization where contact rate parameters are sampled from an estimated distribution each time step. We tested realizations of the models for consistency with the actual SARS data. We discovered that standard disease models with adaptive contact rate parameters still fail to exhibit properties of the real data and therefore other models with better contact descriptions such as small world networks are more suitable for modelling disease transmission.

Acknowledgments

This work was supported thanks to a Hong Kong Polytechnic University grant no. G-YE08.

References


Deterministic Chaos of Human Cardiac Signals

Yi Zhao† and Michael Small†

†Department of Electronic and Information Engineering, Hong Kong Polytechnic University
Hong Kong, China
Email: zhao.yi@polyu.edu.hk, ensmall@polyu.edu.hk

Abstract—We employ a pseudo-periodic surrogate (PPS) algorithm to identify the presence of deterministic chaos in healthy human cardiac time series. We also perform prediction analysis from ECG to pulse data to confirm that the two kinds of signals conform to deterministic processes. Experimental results of on seven healthy subjects indicate that ECG and pulse data are both inconsistent with a noise driven periodic orbit, and bounded aperiodic determinism exists in both of them.

1. Introduction

Whether or not the human cardiac system is chaotic has long been a subject of interest in biomedical application of nonlinear timeseries analysis. Estimation of dynamic invariants from timeseries [1] [2] reveals that electrocardiogram (ECG) signals for normal and pathological cases have a finite non-integer correlation dimension and positive Lyapunov exponent, and indicate the possibility of deterministic chaos.

Notably, linear surrogate data methods have been proposed in recent years to test for the presence of determinism in ECG data [3] [4]. Linear surrogate data methods are employed to test an observed time series against the hypotheses of [5]: NH0: independent and identically distributed (i.i.d.) noise; NH1: linearly filtered noise; and NH2: a monotonic nonlinear transformation of linearly filtered noise. The three hypotheses are all different forms of linear noise processes (albeit possibly subject to static nonlinear filter). Based in the rejection of these hypotheses we can conclude that ECG data are not linear noise but we cannot determine whether they are deterministic pseudo-periodic chaos or a noisy periodic waveform, both of which are plausible. It is obvious that ECG data usually exhibit strong periodicities and definite nonlinearity. It is natural to not only ask whether the ECG data of the healthy human are deterministic or consistent with a stochastic process, but to further wonder whether they are consistent with pseudo-periodic deterministic chaos or a noise driven periodic orbit.

Relatively few attempts to identify the presence of determinism in blood pressure propagation (pulse data) are found in the literature. But pulse pressure measurement (“feeling the pulse” 1) has been popular diagnosis in traditional Chinese medicine (TCM) for centuries. So it is also significant to examine whether pulse signals are deterministic (and possibly confirm to the deterministic origin of ECG data).

So in the current work we significantly extend the previous works [3] [4] to test for the presence of pseudo-periodic determinism in both normal ECG and pulse data. We apply the pseudo-periodic surrogate data method, capable of distinguishing between noise driven periodic orbit and pseudo-periodic chaos, to seven healthy subjects’ ECG and pulse data.

For two deterministic signals, even short-term prediction from one to the other would be impossible if one or both are independent or stochastic, but it is possible to make the prediction from one to another if they exhibit deterministic components and if these deterministic components are related. We therefore adopt neural networks with adequate generalization, which follow the methods proposed and validated in [6] [7], to test predictability between ECG and pulse data for each case.

2. Pseudo-periodic Surrogate Algorithm

The rationale of the surrogate data method is that for a given hypotheses one generates an ensemble of artificial surrogate data by using a corresponding surrogate randomization algorithm which ensures that the generated surrogate data are consistent with that hypothesis. One then applies some test statistics to both the surrogate data and the original data.

Referring to the three commonly employed null hypotheses described previously, three corresponding algorithms to generate surrogate data are known as Algorithm 0, Algorithm 1, and Algorithm 2. For the sake of brevity we do not describe them in detail. The PPS algorithm we employ in this paper provides an entirely new surrogate generation algorithm that tests the null hypotheses that an observed timeseries is consistent with a stable periodic orbit with dynamic noise (i.e. noise-driven periodic orbit). This algorithm [8] can be described as follows:

The observed time series are converted to produce a vector time series \( \{ z_t \} \) representative of the underlying dynamics in phase space [9]. The surrogate data are then generally put three fingers (like three sensors) on the patient’s wrist. They do eventually feel the blood pressure propagation.

---

1When feeling pulse, traditional Chinese medicine practitioners usually put three fingers (like three sensors) on the patient’s wrist. They do eventually feel the blood pressure propagation.
ated by randomly selecting an initial point from \( z_i \). Successive points are generated by choosing a neighbor \( z_j \) of the currently selected point \( z_i \) with the certain noise radius, \( \rho \). The noise radius determines the deviation of the neighbor from the original one. For examples, for large \( \rho \) the selected neighbor \( z_j \) is far away from the current point \( z_i \); for small \( \rho \) the selected neighbor \( z_j \) is close to the point \( z_i \). That successor \( z_{i+1} \) of that neighbor becomes the new current point and we continue selecting neighbors until completing the vector time series.

By tuning \( \rho \) we may generate surrogate data that preserve all the underlying dynamics of the data. Alternatively, we may select neighbors so that the general periodic dynamics are preserved and others are destroyed. In this case surrogate data address the hypothesis of a noise driven periodic orbit. We use the latter scenario to test whether human cardiac signals (ECG and pulse data) be consistent with a noise driven periodic orbit.

For the choice of test statistics, we employ algorithmic complexity [10], rather than the more popular correlation dimension. Correlation dimension is sensitive to noise, even very small noise [11], which makes its application limited in field measurements. Complexity aims to measure the regularity of the finite specified sequence, which behaves differently for deterministic and for random sequences. Compared to existing techniques, algorithmic complexity has the great advantage of small computational cost and is well suited for real-time implementation. To compute this quantity from a time series, one first needs to convert a time series to a symbolic sequence. We do this by converting each point of the time series to one of three symbols, 0, 1, and 2 with equal probability. We then apply the Lempel-Ziv algorithm [10] to compute the number of unique symbol sequences. Finally the calculated number is normalized by the number of unique symbols for a random sequence. A normalized complexity near zero denotes that the time series is deterministic and periodic (a stable fixed point without noise). Complexity near one means that the signals is highly random. Complexity of chaotic time series is between zero and one.

3. Application of PPS method to cardiac signals

ECG and pulse data of healthy subjects (2 female and 5 male volunteers) are measured in quiet and relaxed situation to minimize disturbances. The measurement device is PowerLab 4/25 of ADInstruments. The sampling rate is set to 100 Hz with the resolution of 16 bits. We measure each volunteer’s surface ECG data and pulse data on the fingertips simultaneously. The amplitude of ECG data (\( \mu V \)) is greatly different from that of pulse data (\( mV \)) so we normalized the measurement data. By using the PPS method we generate 50 surrogate data for ECG (1000 points) and pulse data (1000 points) of each volunteer and then compute complexity of the original data and surrogate data. Figure 1 presents typical results of ECG and pulse data for one subject.

![Figure 1](image)

**Figure 1:** The histograms show the distribution of complexity of 50 surrogates for ECG (a) and pulse data (b). The x-axis of the solid line is the complexity of the ECG (a) and pulse data (b).

In Figure 1 complexity of original ECG and pulse data are both far away from distributions of complexity of their surrogate data. We therefore can reject the hypothesis that ECG and pulse data are noise driven periodic orbit, i.e. the dynamics of ECG and pulse data both are distinct from a noise driven periodic orbit. This result indicates that pseudo-periodic deterministic chaos may exist in both of them.

As a comparison we demonstrate the application of this algorithm with same value of the parameter \( \rho \) to normalized periodic sine and chaotic Rössler data. The period of the sine data is the same as that of the preceding data. Results for the sine timeseries (1000 points) and the chaotic Rössler data (1000 points) are illustrated in Figure 2.

According to results of comparative experiments (Figure 2) we find that the Rössler data are not noise driven periodic orbits but the sine data are consistent with this hypotheses. It means that the PPS algorithm can make a correct decision on the hypotheses of noise driven periodic orbit. More significantly, experimental results of ECG and pulse data are consistent with results of the chaotic Rössler data and distinct from that of the periodic sine data. We therefore conclude that pseudo-periodic deterministic chaos exists in both ECG and pulse data. Moreover, this conclusion taken on cardiac data of this subject is consistent with all the other subjects. Table 1 summarizes results of application of the PPS method to ECG and pulse data of all the seven

---

![Table 1](image)

**Table 1:** Summary of results of application of the PPS method to ECG and pulse data of all the seven volunteers.
4. Nonlinear Modeling from ECG to Pulse Data

In this section we employ neural networks to perform prediction from ECG to pulse data and vice versa so as to model the relationship between them. However, overfitting is a common phenomenon in neural networks with a fairly large number of neurons. So neural networks with different numbers of neurons are used to make prediction between ECG and pulse data. We then employ minimum description length (MDL) to determine the optimal neural network, which are observed to provide adequate generalization [6] [7].

Firstly we perform the prediction from ECG to pulse data. The 3600 data points are selected to train the neural networks with another 800 points as the testing set. For all the seven subjects the testing pulse data have high correlation with the corresponding prediction of the optimal neural network, as illustrated in Table 2.

Finally, we apply surrogate data method to test whether the prediction error are just i.i.d. noise, i.e. to ensure the ECG data can exactly predict the pulse data. The surrogate data method we employ is the small shuffled surrogate data method [12], which tests against the hypotheses of independently distributed noise. Since in the residual relatively large amplitudes periodically appear it is obvious that this residual is not i.i.d. noise and the surrogate data method with the hypotheses of NH0 is therefore not applicable. But we are not sure whether this residual is consistent with the independently distributed noise (another random noise) so we apply the small shuffled surrogate method to this model residual. Figure 3 shows results of application of small shuffle surrogate to the residual of the optimal neural networks for Subject 1.

Referring to Figure 3 we demonstrate that this residual is random noise and thus ECG data can precisely predict pulse data by optimal neural networks estimated by MDL. Consequently, we conclude that ECG and pulse data conform to deterministic processes since there is good short-term prediction from ECG to pulse data.

However, we still need to test the causal relationship between ECG and pulse data. When we repeat the previous procedures to perform prediction from pulse data to ECG data. The prediction results are rather poor, that is, the prediction error contains deterministic components, even if we try more training data, longer training iteration time, and different training algorithms. We further conclude that the ECG and pulse data do not conform to the same deterministic process, and ECG data comprise certain deterministic information, which the pulse data can not replicate or do not contain. So pulse measurement does not contain complete information of the human cardiovascular system. This may provide theoretical evidence that auscultation and olfaction, interrogation, and palpation are also necessary ways of diagnosis in traditional Chinese medicine.\(^2\)

\(^2\)Diagnosis-observation (of patients’ complexion, tongue, and etc.), auscultation and olfaction, interrogation, feeling pulse constitute the fundamental diagnosis in TCM.
Table 1: Results of application of the PPS method to ECG and pulse data for all subjects.

<table>
<thead>
<tr>
<th>Subject</th>
<th>ECG data</th>
<th>Pulse data</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.369</td>
<td>0.346</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.442</td>
<td>0.416</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.421</td>
<td>0.365</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.496</td>
<td>0.509</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.421</td>
<td>0.409</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.553</td>
<td>0.409</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.421</td>
<td>0.358</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.553</td>
<td>0.321</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.421</td>
<td>0.321</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.553</td>
<td>0.282</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.421</td>
<td>0.282</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.553</td>
<td>0.314</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
<td></td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.421</td>
<td>0.314</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.553</td>
<td>0.389</td>
</tr>
</tbody>
</table>

Table 2: Cross coefficient of the pulse data and corresponding prediction from ECG to it for all subjects.

<table>
<thead>
<tr>
<th>Subject</th>
<th>Cross coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.997</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.996</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.993</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.992</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.995</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.993</td>
</tr>
<tr>
<td>c</td>
<td>&lt; c &gt;</td>
</tr>
<tr>
<td>c&lt;sub&gt;max&lt;/sub&gt;</td>
<td>0.993</td>
</tr>
<tr>
<td>c&lt;sub&gt;min&lt;/sub&gt;</td>
<td>0.990</td>
</tr>
</tbody>
</table>

5. Conclusion

We demonstrate a variety of tests to seven healthy human ECG and pulse data so as to identify the presence of deterministic chaos in these cardiac data. Experimental results of application of the PPS method illustrate both ECG and pulse data are not consistent with noisy periodic orbits and pseudo-periodic deterministic chaos may exist in them. For comparison, we also demonstrate the same method to periodic sine data and chaotic R"ossler data. The results of ECG and pulse data are consistent with the results of R"ossler data but different from that of sine data. This provide more evidence that human cardiac data are consistent with pseudo-periodic determinism.

By prediction analysis between ECG and pulse data we notice that ECG and pulse data both stem from deterministic processes but do not exactly conform to the same deterministic process. These results also suggest the utility of nonlinear methods in medical diagnosis and potential future treatment regimes (traditional Chinese medicine).

Acknowledgment

This work was supported by Hong Kong University Research Council Grant (PolyU 5235/03E).

References


Design of a Low EMI Hysteretic Current-Controlled DC/DC Boost Converter
Via Chaotic Perturbation

Michele Balestra†, Fabio Pareschi‡, Gianluca Setti† and Riccardo Rovatti‡

†School of Engineering, University of Ferrara
Via Saragat 1, Ferrara 44100, Italy
‡ARCES, DEIS, University of Bologna, Via Toffano 2/2, Bologna 40135, Italy
Email: mbalestra@ing.unife.it, gsetti@ing.unife.it, rrovatti@arces.unibo.it

Abstract—In this paper we aim to design a self-oscillating DC/DC boost converter operating in continuous-conduction mode using an hysteretic control, using a chaos-based signal processing approach to improve EMC of the converter. We analyze the effectiveness of this technique both by computing the Power Density Spectrum of the inductor current and by experimental results. We show that the proposed chaos-based scheme can be chosen to greatly enhance the EMC of the systems with no degradation in the converter performance.

1. Introduction
The goal of reducing Electro-Magnetic Interference (EMI) emissions is of great importance. In the last years increasing attention has been paid to EMI reduction by exploiting nonlinear dynamics and signal processing techniques. This is mainly done by altering conventional power converter structure, by introducing complex non-periodic or even chaotic perturbation in the periodic control signals. Hence, periodicity of currents and voltages is disrupted so that the energy of the interfering signal that was originally concentrated at the harmonic frequency is spread over a large bandwidth. This obviously reduces the peak value of Power Density Spectrum (PDS) of any interfering signal emitted by the system [8].

These perturbations may be due to the operation of switching circuits in a chaos-like regime (see e.g. [4][2][5][6] and reference therein) or to the explicit coupling of the system with a periodic, random or chaotic signal generator (see e.g. [1][7], [3][9][10][14] and references therein). In particular, in [4][2][5][13] the chaotic mode of operation of the power converter was proposed as a way to broaden the PDS of the converters currents and voltages.

Although a significant PDS peak value reduction can be achieved with respect to converters operating in periodic regime [5], the spread in the signals PDS provokes an increase in the output voltage ripple, consequently degrading the system performance. In [6] a thorough study is reported on the PDS of inductor current of a chaotically operating boost converter that suggests a way to further reduce EM coupling and interference.

The other class of techniques [1][7][3][14] is also very effective in reducing the EMI problem at its origin. For example [7][14], a reduction of up to 20dB in the peak power spectral density with respect to other patented method can be achieved via a chaos-based frequency modulation (FM) of the timing signal driving an open-loop converter.

The goal of this paper is to design a self-oscillating DC/DC boost converter operating in continuous-conduction mode using an hysteretic control. This structure simplify both design and converter complexity since it don’t require an external clock, while improving the output ripple performance for high voltage conversion ratios as chaos-like regime cannot take place. We analytically design a non-linear perturbation to be applied on the hysteretic control loop, exploiting the knowledge on chaos-based EMI reduction techniques [7][9][10][12][14], and to assess the impact of this new methodology on the converter EMC performance. Finally, we will also develop a prototype to experimentally verify the validity of our approach.

2. Hysteretic current-controlled DC/DC boost converter

Figure 1: Hysteritic current-controlled DC/DC Boost converter scheme

The current-controlled boost converter is shown in Figure 1, and it is assumed to operate in continuous-conduction mode. A voltage feedback loop provide the reference current $I_{ref}$, which is set to be the hysteresis window midpoint value, while the hysteresis threshold control provide the current levels which limit the excursion of the inductor current. When the switch is closed, the inductor current $i_L$ increases until the high threshold $I_{th}^{h}$, then the switch opens and the current falls, reaching the low threshold $I_{th}^{l}$ causing switch to close again, starting a new cycle. We assume that the output voltage $V_{out}$ is essentially constant and that the switch ON and OFF time intervals are short enough so that the current can be assumed to be piecewise linear (see figure 2).

Unlike other family of current-controlled DC/DC boost converter, which can exhibit an intrinsic chaotic behav-
ior when the output voltage is at least twice the input voltage[4], the hysteretic current-controlled boost converter with fixed threshold values has a periodic regime, which makes it suitable to implement any voltage conversion ratio without the need of any strategies to avoid unwanted behaviors.

If we indicate by $m_1 = V_1/L$ and $m_2 = (V_{out} - V_0)/L$ the slope magnitude of the inductor current when the switch is respectively ON or OFF, we have

$$f_k = \frac{m_1 + m_2}{2m_1m_2(f_{th} - f_{th}')}, \quad (1)$$

where $f_k$ is the actual inductor current frequency induced by the thresholds and $f_{th} = I_{ref} - I_r, f_{th}' = I_{ref} + I_r - \xi.$

When the voltage feedback is introduced, the reference current $I_{ref}$ is controlled by means of the output voltage $V_{out}$. As it is commonly assumed in the analysis and design of this class of converters, we suppose that the voltage feedback outer loop is characterized by a much slower dynamics compared to the one of the current control, so that $I_{ref}$ can be considered as almost constant during our analysis.

3. PDS Estimation

Figure 2 shows a possible inductor current waveform. The value of $I_r$ set the amplitude of the hysteresis window, thus affecting both the output voltage ripple and the operating frequency of the switch. If the converter is working in a periodic regime (i.e. $\xi(t) = 0$) the PDS of the inductor current is concentrated at the harmonic frequencies of the switching timing signal given by (1). To control the harmonic frequencies position we need to act on the hysteresis window amplitude. To do this, we introduce a noise-like PAM signal

$$\xi(t) = A \sum_{k=0}^{\infty} x_k g(t - kT), \quad (2)$$

where the symbols $\{x_k\}$ are assumed to be generated by a one-dimensional chaotic map, with $x_{k+1} = M(x_k), M : [0, 1] \mapsto [0, 1], T$ is the symbols update period, $A$ is the amplitude of the signal and $g(\cdot)$ is a function which is 1 in $[0, T]$ and 0 elsewhere.

To analyze the effect of the chaotic perturbation on the converter behavior and to evaluate its impact on EMI reduction, we need to compute the PDS of $i_L(t)$. We will assume that $T$ is large enough to neglect the transient effect in the value of $i_L(t)$ occurring after a change of the modulating symbol from $x_k$ to $x_{k+1}$.

With this, we can assume that the converter evolves through a sequence of steady-state periodic behaviors with different timing signal frequencies (and thus current and voltages) and, in particular, that the inductor current reaches the new stable periodic solution in a negligible time after each symbol change. Exploiting these hypothesis, we can easily write

$$i_L(t) = \sum_{k=-\infty}^{\infty} s_k(t) g_T(t - kT), \quad (3)$$

where $s_k(t)$ indicates the periodic triangular wave of the kind shown in figure 2, extended on the whole real axis, and corresponding to the steady state behavior of a converter working in periodic mode at frequency $f_k$. Since $M$ is mixing, $s_k(t)$ is cyclo-stationary [11][12] so that its PSD $\Phi(f)$ is well defined and can be expressed as

$$\phi(f) = \lim_{MT \to \infty} \frac{1}{2MT} E \left[ |S_M(f)|^2 \right], \quad (4)$$

where

$$S_M(f) = \int_0^{MT} \dot{i}_L(t) e^{-i2\pi f t} d\tau \quad \text{and where } \int_{\cdot}^{\cdot} \text{ is the expectation with respect to the process generating the modulating sequence. By defining}$$

$$\sigma_n(f) = e^{-i2\pi fnT} \int_0^T \dot{i}_L(t) (\xi + nT) e^{-i2\pi f t} d\xi \quad (5)$$

we can write $S_M(f) = \sum_{n=0}^{M-1} \sigma_n(f)$, so that we obtain [15]

$$|S_M(f)|^2 = \sum_{n=0}^{M-1} \sum_{n=0}^{M-1} \sigma_n(f) \sigma^*_n(f) = \sum_{n=0}^{M-1} \sigma_n(f) \sigma^*_n(f) + 2 \text{Re} \left[ \sum_{n=0}^{M-1} \sum_{k=1}^{M-1} \sigma_{n+k}(f) \sigma^*_n(f) \right] \quad (6)$$

where $\cdot^*$ denotes complex conjugation. Hence, (4) can be written as

$$\Phi(f) = K_1(f) + 2 \text{Re} \left[ K_2(f) \right] \quad (7)$$

Exploiting the periodicity of $\sigma_n(t)$ we have

$$\sigma_n(f) = e^{-i2\pi fnT} F \left[ \sum_{m=0}^{\infty} \cos(2\pi m f n t + \varphi_m) \right] \quad (9)$$

where $F(\cdot)$ is the Fourier Transform operator, so that

$$\sigma_n(f) = e^{-i2\pi fnT} \sum_{m=0}^{\infty} A^m_n e^{i\varphi_n} G_T(f - m f n) \quad (10)$$

where $G_T(f) = F[g_T(t)].$

The explicit calculation of (4) can be a hard task, but we can still obtain some useful information on the actual PDS. From now on we will assume that the $\{x_k\}$ are iid random variables. With this, we can write

$$\lim_{T \to \infty} E \left[ e^{i\varphi_n - i\varphi_0} G_T(f - m f n) G_T(f - m f n + k) \right] = 0 \quad (11)$$
as \( G_T(f - f_0) \) is vanishing to the limit for every \( f \neq f_0 \). It follows that

\[
\lim_{T \to \infty} K_2(f) = 0
\]  

(12)

Similarly, we can write

\[
K_1(f) = \frac{1}{2T} \sum_{m=0}^{\infty} E \left[ \frac{A_m^2 G_T(f - \frac{m}{c(I_r + A)})}{m} \right]^2
\]  

where \( c = 2m_1m_2/(m_1 + m_2) \). Assuming that the perturbation provided by the symbols \( \{x_n\} \) on the amplitudes \( A_m \) is negligible (i.e. \( A_m \approx \) constant), we can write:

\[
\lim_{T \to \infty} \Phi(f) = \lim_{T \to \infty} K_1(f) = \sum_{m=1}^{\infty} \frac{m\lambda_m}{2\lambda} \rho\left(\frac{m}{cA} - \frac{I_r}{A}\right) \frac{1}{f^2}
\]  

(14)

where \( \rho(x) \) is the probability density function (PDF) of the symbols \( \{x_n\} \) i.e. the invariant density of the chaotic map \( M_0(x) \).

Equation (14) shows that \( \Phi(f) \) is a superposition of spectra around each harmonic of the \( i_L(t) \) in a converter working in the periodic regime. These spectra have the same shape of the PDF of the symbols multiplied by the distortion factor \( 1/f^2 \), hence the maximum spread spectrum spread and therefore maximum peak reduction strictly relies on the choice of the symbol generator. This result is still a good approximation of the PDS trend even when the symbol update period \( T \) is kept large enough compared to the switching period.

Exploiting the results in [14] that assures that we can substitute chaotic modulating symbols for the iid random ones if \( M \) has a sufficiently strong mixing behavior, we can design a chaotic map with the proper PDF in order to obtain a quite flat PDS shape.

A chaotic map \( M_0(x) \) suitable to this aim is defined by:

\[
M_0(x) = \begin{cases} 
1 - 2\left(\sqrt{2(1-x)} - 1\right)^2 & -1 < x < 0.5 \\
-3 + 4x & 0.5 < x \leq 1 
\end{cases}
\]  

(15)

The map is depicted in Figure 3 along with its invariant density.

4. Experimental results

A converter prototype was realized, with \( L = 1.5 \)mH, \( C = 100 \)µF, \( R = 1002 \) and employing a NBR1540 power diode, a IRF540 power MOSFET for the switch and \( V_{in} = 6 \)V. The center band frequency was set at \( f_0 = 12 \)kHz.

The chaotic PAM was implemented using a TMS320C6711 DSP board, with \( T = 0.5 \)ms.

We compare here the results obtained in the periodic regime and when the perturbation is applied.

![Figure 3: The chaotic map \( M_0(x) \) (left) and its invariant density (right).](image)

![Figure 4: PDS of the inductor current in the periodic regime case.](image)

![Figure 5: PDS of the inductor current in the chaotic case using a Bernoulli map as symbols generator.](image)

First we assume \( V_{out} = 14 \)V with no perturbation added to the threshold levels, so that the converter operates in the classical periodic setting. This is our reference r-case. Then we consider the chaotic driven case ch-case, where \( \xi(t) \) provided by (2) is non-null, the chaotic symbols were generated by a Bernoulli shift and the maximum frequency deviation is set to \( \Delta f = 3.5 \)kHz.

Figure 4 shows the measured PDS of \( i_L \) in the r-case. Its peak value is \( \Phi_{\text{max}} = 4.50 \text{dBm} \) while the output voltage ripple is \( R_{\text{out}} = 198 \)mV.

Figure 5 shows the trend of the PDS in the ch-case. As expected, the introduction of the chaotic perturbation is very effective in reducing EMI since \( \Phi_{\text{max}}^{ch} = -3.21 \text{dBm} \) with a ripple \( R_{\text{out}}^{ch} = 199 \)mV, i.e. practically unchanged with respect to the r-case.

As the chaotic map selected has an uniform invariant density, the PDF of the symbols is constant. Figure 5 clearly shows the impact of the distortion term \( 1/f^2 \) in (14), as the spectrum has a quite linear trend in the logarithmic scale. This behavior is not optimal for EMI reduction. To achieve a flatter spectrum, we introduce the chaotic map \( M_0(x) \) defined by (15).

The map \( M_0(x) \) was implemented by means of a DSP and the resulting inductor current PDS is shown in Figure 6. The PDS has a quite constant shape around each harmonic, showing a good agreement with the theoretical
prediction. Its peak value is $\Phi_{\text{ch-fm}} = -4.39$dBm while the output voltage ripple is $R_{\text{vin}}$ = 198mV, so still unchanged by the perturbation. The improvement with respect to the simple Bernoulli driven perturbation is about 1.18dB which shows the great impact of such PDS optimization.

5. Conclusions

In this paper we studied the effect of a suitable chaos-based perturbation applied to threshold levels of a hysteretic current-programmed DC/DC boost converter with voltage feedback loop. This perturbation on the comparators thresholds reflects in reiterated frequency changes of the main switch control function, so disrupting any periodicity and producing a broaden power density spectrum of the inductor current, thus reducing the peak PDS value.

We provided an analytical expression to approximate the inductor current PDS, and we verified its validity by means of measures performed on a converter prototype. Experimental results showed that the spreading doesn’t alter the output voltage ripple, while $i_L$ lowers the PDS peak value and increases the circuit EMC.

However, an uniform PDF symbols source doesn’t give the lower PDS peak, as the spectrum shape is affected by a non linear distortion resulting in a non-flat, thus non-optimized, PDS shape.

We provide a simple chaotic map which has an invariant density suitable to optimize the EMI performances. Measures show that the resulting spectrum shape is quite flat so that the behaviour with respect to EMI is further improved.

Compared to the techniques in [7][9][10] this approach doesn’t require a frequency modulator for the PDS spreading to take place, so the implementation of the control circuit is further simplified.

The symmetry structure of the perturbation application is effective in keeping a low drift in the output voltage fluctuation, as the inductor current mean value remains quite unaltered through the different hysteresis window amplitudes.

References


Figure 6: PDS of the inductor current in the chaotic case using Mo(x) as the symbol generator.
A digital spiking neuron and its simple learning

Hiroyuki Torikai†, Yoshiaki Shimizu‡ and Toshimichi Saito†

†EECE Dept., Hosei University, Koganei, Tokyo 184-8582, Japan
‡Tokyo Metropolitan Government

Abstract—A digital spiking neuron can generate spike-trains having various spike intervals. Some basic relations between parameters of the neuron and characteristics of the spike-train are clarified. Using the relation, a basic learning algorithm of the DSN is proposed in order to search parameter values that are suitable for some applications. Although search space is extremely large, the algorithm can find suitable parameter values in a realistic time.

1. Introduction

We study a digital spiking neuron (DSN) which is an ultra-discrete system, and its dynamics is inspired by an analog spiking neuron model [1]-[3]. Fig.1(a) shows basic dynamics of a spiking neuron. A potential $v$ usually increases by integrating a DC stimulation. When the state $v$ reaches a threshold $\theta$, it is reset to a periodic base signal $\beta(t)$ and a firing spike $Y=1$ is generated. Repeating such integrate-and-fire behavior, the neuron can generate periodic and non-periodic spike-trains [4][5]. Fig.2(b) shows basic dynamics of the DSN. A black box represents a digital state (corresponding to the potential $v$) and is usually shifted upward. When the black box reaches the highest position (corresponding to the threshold $\theta$), it is reset to a white circle (corresponding to the base signal $\beta(t)$) and a firing spike $Y=1$ is generated. Repeating such shift-and-fire behavior, the DSN generates a spike-train $Y$. The DSN consists of digital registers and wirings among them. The wiring pattern is a crucial parameter because it determines shape of a map which governs dynamics of spike interval [1]-[3]. That is, the DSN can realize various dynamics of spike-train by adjusting the wiring pattern. Now, one of fundamental problems is: how we can synthesize a DSN that generates spike-train having desired characteristics.

In this paper we investigate one approach toward such a problem, based on an understandable example application. First we analyze some characteristics of the spike-train, e.g., period and number of spikes. We then clarify a basic relation between the wiring pattern and characteristics of the spike-train. Using the relation, we can propose a simple learning algorithm in order to search wiring patterns of the DSN that are suited for ranging and positioning applications. We show that the algorithm can find suitable wiring patterns in a realistic time. It should be noted that carpet-bombing search is not realistic because the number of possible wiring patterns increases exponentially as the number of registers increases.

Significances and motivations for studying the DSN are many, including the following points. (1) Spiking neurons and their networks can exhibit rich phenomena (e.g., synchronization clusters of spike-trains) that relate to engineering applications [5]-[8]. Ultra-discrete systems can also exhibit rich phenomena, e.g., beautiful spatio-temporal pattern generation by cellular automata and pseudo-random number generation by shift register generators [9]-[11]. It is fundamentally interesting to study fusion of such important dynamical systems. (2) Spike-based systems have some engineering advantages, e.g., fast computation of pulse-coupled neural networks [7][8]; and low power consumption and multiplex communication capability of impulse communications [12][13]. Ultra-discrete systems can be implemented easily and systematically by reconfigurable hardware such as FPGA. The DSN may be a prototype in order to fuse advantages of both systems. (3) The wiring pattern of the DSN can be adjusted dynamically by FPGA. Hence the presented learning algorithm may be fundamental to develop on-chip learnings of the DSN and/or its pulse-coupled network.

2. Digital Spiking Neuron

In this section we introduce the digital spiking neuron (DSN) and its basic dynamics [1]. The DSN operates on a discrete time $t = 0, 1, 2, \cdots$. Fig.2(a) shows an example of the DSN. The DSN has $M$ pieces of $p$-cells indexed by $i \in [0, 1, \cdots, M-1]$, where $M \geq 2$. Each $p$-cell has a digital state $p_i(t) \in \{0, 1\} \equiv B$, where "$\equiv$" denotes "definition" hereafter. The $p$-cells are ring-coupled and governed by

$$p_{i+1 \mod M}(t+1) = p_i(t).$$

(1)

In this paper initial states of the $p$-cells are fixed: $p_0(0) = 1$ and $p_k(0) = 0$ for all $k \neq 0$. Then the $p$-cells oscillate with period $M$. As shown in Fig.2(a), the DSN has
Reconfigurable wirings
\[ p_{cells} \times x_{cells} \]

Basic dynamics of the \( x \)-cells is shown in Fig.2(b). If the black box (that represents \( x_j = 1 \)) is below the highest \( x \)-cell (that is indexed by \( N - 1 \)), it is shifted upward. If the black box reaches the highest \( x \)-cell at \( t = t_0 \) (i.e., \( x_{N-1}(t_0) = 1 \)), the DSN fires: the black box at \( t = t_0 + 1 \) is reset to the position of the white circle (that represents \( b_j = 1 \)) at \( t = t_0 \), and the DSN generates a spike \( Y(t_0) = x_{N-1}(t_0) = 1 \). Repeating such shift-and-fire dynamics, the DSN generates a spike-train

\[ Y(t) \equiv x_{N-1}(t), \quad t = 0, 1, 2, \ldots \]  

The shift-and-fire dynamics may be regarded as digital version of integrate-and-fire dynamics of an analog spiking neuron [5]. The upward shift of the digital state \( X(t) \) corresponds to integration dynamics of an analog membrane potential. The reset of \( X(t) \) and generation of a spike \( Y(t) = 1 \) correspond to firing dynamics of the analog spiking neuron. Hence we refer to the circuit in Fig.2(a) as the digital spiking neuron.

As a result, the DSN is governed by the set of Equations (1), (3), (4) and (5). Also the DSN is characterized by the parameters \( M, N \) and \( A \). The number of possible wiring matrices \( A \) is

\[ \# \text{ wiring matrixes} = N^M. \]  

The DSN has a controllable initial state vector \( X(0) = (x_0(0), \ldots, x_{N-1}(0))^T \) of the \( x \)-cells. In this paper we assume that only one element in \( X(0) \) is 1.

The DSN generates a periodic spike-train in a steady state because it has the finite states \( \langle P, X \rangle \) and the discrete time \( t \). The DSN may have multiple co-existing periodic spike-trains and generates one of them depending on the initial state \( X(0) \). Hence the DSN may exhibit complicated transient phenomena to multiple co-existing periodic spike-trains [1]. In this paper we focus on spike-trains having long periods and do not deeply discuss co-existence and transient phenomena.

3. Application example and Basic learning

In this section we clarify basic relation between the wiring matrix \( A \) and characteristics of the spike-train \( Y \). Based on the relation, we also propose a basic learning algorithm for some applications. For simplicity, let us focus on the case of \( M = N \). An integer \( T \) is said to be period of a spike-train \( Y(t) \) if \( T \) is the minimum multiple integer of \( M \) such that \( Y(t + T) = Y(t) \) for all \( t \geq 0 \). The number \( Q \) of spikes during the period \( 0 \leq t < T \) is said to be spike-number. The spike-train \( Y(t) \) in Fig.2(b) has period \( T = 21 \) and spike-number \( Q = 7 \). Let us define an auto-correlation function of a spike-train \( Y(t) \) as follows:

\[ C_Y(\tau) \equiv \langle \sum_{i=0}^{T-1} Y(t)Y(t+\tau) \rangle. \]  

Fig.2(c) shows the auto-correlation function \( C_Y(\tau) \) of the spike-train \( Y(t) \) in Fig.2(b). Let \( R \equiv \max_{1 \leq \tau \leq T-1} |C_Y(\tau)| \)
Mutation rule: Let $G$ be an $M \times M$ matrix whose elements are given by $g(j, i) = a(M - j + i \pmod{M})$ for all $i$ and $j$. Let $\alpha \in \{0, 1, \ldots, M - 1\}$ and $\beta \in \{0, 1, \ldots, M - 1\}$ be arbitrary integers. Let us exchange the $\alpha$-th and the $\beta$-th columns of $G$, and also exchange the $\alpha$-th and the $\beta$-th rows. The resulting matrix is denoted by $G'$. Let $A'$ be an $M \times M$ matrix whose elements are given by $a'(j, i) = g(M - j + i \pmod{M})$ for all $i$ and $j$. $A'$ is referred to as a mutated wiring matrix of $A$, where $\alpha$ and $\beta$ are referred to as mutation positions.

For example, if we mutate the wiring matrix $A$ in Equation (2) with mutation positions $(\alpha, \beta) = (1, 2)$, we obtain

$$A' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

where "′" denotes "mutated version" hereafter. Fig.3(a) shows the DSN having the mutated wiring matrix $A'$ in Equation (8). This DSN generates a spike-train $Y'(t)$ in Fig.3(b) that is characterized by $(T', Q', R') = (28, 7, 4/7)$. We can confirm that the spike-trains $Y(t)$ in Fig.2(b) and $Y'(t)$ in Fig.3(b) have the same spike-number $T = T' = 7$. Actually, we have the following conjecture.

Conjecture 1: Suppose a DSN with a wiring matrix $A$ generates a spike-train $Y(t)$ with spike-number $Q = M$. Then a DSN with a mutated wiring matrix $A'$ generates a spike-train $Y'(t)$ with spike-number $Q' = M$. $Y(t)$ and $Y'(t)$ may have different period (i.e., $T \neq T'$) and 2nd peak of auto-correlation function (i.e., $R \neq R'$).

Based on this conjecture, we can propose a basic learning algorithm of the wiring matrix $A$ for preferable cost functions C1-C3.

Learning algorithm

Step 0. Initialize the wiring matrix as follows: $a(M - 1, i) = 1$ for all $i$; and $a(i, i) = 0$ for all $i$ and $j \neq M - 1$. If the DSN generates a spike-train $Y(t)$ characterized by $(T, Q, R) = (M, M, 1)$ as shown in Fig.4(a). We note that the spike-train $Y(t)$ is to have the possible maximum spike-number $Q = M$. Initialize a counter $k$ of learning iteration number to 0.

Step 2. Using two random integers $\alpha \in \{0, 1, \ldots, M - 1\}$ and $\beta \in \{0, 1, \ldots, M - 1\}$, mutate the wiring matrix $A$ into $A'$. We note that Conjecture 1 says that any mutated DSN generates a spike-train $Y'(t)$ having the possible maximum spike-number $Q' = M$. If $Y'(t)$ has lower 2nd peak $R'$ of auto-correlation function than that of $Y(t)$ (i.e., $R \leq R'$), go to Step 3. Otherwise, go to Step 4.

Step 3. Update the wiring matrix as $A = A'$. Go to Step 4.

Step 4. Increment the counter as $k = k + 1$. Let $K$ be a given maximum iteration number. If $k \leq K$, go to Step 2. Otherwise, terminate the learning algorithm.
Fig. 4 shows basic learning dynamics for $M = 20$. We can confirm that the period $T$ increases and the 2nd peak $R$ of the auto-correlation function decreases as the learning continues. That is, preferable cont functions C1 and C3 can be obtained. We can also confirm that the spike-number $Q$ always takes the possible maximum value 20, i.e., the cost function C2 is always the most preferable one. Fig. 5 shows learning characteristics for $M = 20$, where the graphs are averages for 10 different learning results. It seems that the learning almost converges after $k = 100$ iterations. Recall that the number of possible wiring matrixes is $20^{20}$ and then carpet-bombing search is impossible in a realistic time.

4. Conclusions

We have analyzed basic characteristic of the DSN, e.g., period of the spike-train, spike-number, and auto-correlation function of the spike-train. We then derived the conjecture that clarifies basic relation between the wiring matrix and the spike-number. Based on the conjecture, we have proposed the learning algorithm that can find wiring matrixes that are suitable for ranging and positioning applications. Future problems include: (a) detailed analysis of the DSN and the learning algorithm, (b) FPGA implementation of the learning algorithm, and (c) synthesis and analysis of a pulse-coupled neural network of the DSN having some useful application.

References

Probabilistic Controllability Analysis of Piecewise Affine Systems

Shun-ichi Azuma† and Jun-ichi Imura‡
†Kyoto University, Uji, Kyoto 611-0011, JAPAN
‡Tokyo Institute of Technology, 2-12-1, O-okayama, Meguro-ku, Tokyo 152-8552, JAPAN

Email: sazuma@i.kyoto-u.ac.jp, imura@cyb.mei.titech.ac.jp

Abstract—This paper introduces a framework of the probabilistic controllability analysis of piecewise affine systems, which has been recently proposed by the authors. First, based on an algebraic characterization of the controllability, three probabilistic algorithms, which are complementary to each other, are derived for approximately solving the controllability problem in a polynomial time with respect to several parameters. Next, it is shown by some examples that the approach is practical.

1. Introduction

The controllability analysis for hybrid systems is one of the important research topics, while the problem is known to be NP-hard/undecidable [1, 2]. So although several deterministic approaches have been developed so far (see e.g., [1]), the computation amount in these approaches increases exponentially with the control time period and the dimension of the continuous state.

On the other side, an alternative, possible approach to solve the controllability problems in a practically short time is to use a probabilistic method. Thus, the authors have most recently proposed in [3, 4, 5] three approaches using a probabilistic method (e.g., [6]) for approximately verifying the controllability of a class of piecewise affine systems with a probabilistic accuracy: a positive one-sided error algorithm (Algorithm P) based on sampling the discrete state trajectory, a negative one-sided error algorithm (Algorithm N) based on sampling the initial continuous state, and a two-sided error algorithm (Algorithm T) based on sampling both the discrete state trajectory and the initial continuous state. Algorithm P is a polynomial-time algorithm with the control time period and the other parameters except for the dimension of the continuous state, which can determine if the system is controllable or is uncontrollable with a probabilistic accuracy. In contrast, Algorithm N is a polynomial-time algorithm with respect to the dimension of the continuous state and the other parameters except for the control time period, which can determine if the system is controllable with a probabilistic accuracy or is uncontrollable. Further, Algorithm T is a polynomial-time algorithm with respect to all parameters, which can determine if the system is either controllable or uncontrollable with a probabilistic accuracy.

In this paper, through the survey of the above three probabilistic algorithms, it is shown that our framework is very practical for the controllability problem of hybrid systems.

Notation: Let $\mathcal{R}$ and $\mathcal{N}(\mathcal{N}_+)$ denote the real number field and the set of nonnegative (positive) integers, respectively. The $i$-th element of the vector $x$ is expressed as $x_{(i)}$, and the difference set of the sets $X_1$ and $X_2$ is expressed as $X_1 - X_2$. For the measurable set $\mathcal{X}$ and its Hausdorff dimension $d_H$, $\text{vol}(\mathcal{X})$ expresses the $d_H$-dimensional Hausdorff measure, and $\text{card}(I)$ expresses the cardinality of the finite set $I$.

2. Characterization of Controllability of Discrete-Time PWA Systems

This paper considers the discrete-time PWA system

$$\Sigma: \begin{cases} x(k+1) = A_I(k)x(k) + B_I(k)u(k) + a_{I,k} & \text{if } x(k+1) \in S_{I+} \\ I(k+1) = I_k & \text{if } x(k+1) \in S_{I-} \end{cases}$$

(1)

where $x \in \mathcal{R}^n$ is the continuous state, $I \in \mathcal{I}$ is the discrete state (it is sometimes called the mode), $\mathcal{I} := \{0, 1, \ldots, M-1\}$ is the set of the discrete state values, $M \in \mathcal{N}_+$ is the number of the discrete state values, $u \in \mathcal{R}^m$ is the control input, $A_I \in \mathcal{R}^{n \times n}$, $B_I \in \mathcal{R}^{n \times m}$, and $a_I \in \mathcal{R}^n$ are constant matrices for mode $I$, $k \in \mathcal{N}_+$ is the sampling time, and $I_+ \in \mathcal{I}$ is the discrete state value at the $(k+1)$-th sampling time. We call $(I, x) \in \mathcal{I} \times \mathcal{R}^n$ the hybrid state (or simply the state). In addition, $S_I$ denotes the subregion of the continuous state assigned to $I \in \mathcal{I}$, given by

$$S_I := \{ x \in \mathcal{R}^n \mid C_I x + d_I \leq 0, \hat{C}_I x + \hat{d}_I < 0 \}$$

(2)

where $C_I \in \mathcal{R}^{p_I \times n}$, $d_I \in \mathcal{R}^{p_I}$, $\hat{C}_I \in \mathcal{R}^{p_{I+} \times n}$, and $\hat{d}_I \in \mathcal{R}^{p_{I+}}$. For this subregion, it is assumed that $\bigcup_{I,J} S_I = \mathcal{R}^n$ and $S_I \cap S_J = \emptyset$ for every $I, J \in \mathcal{I}$ such that $I \neq J$. This assumption guarantees that $I$ is uniquely determined for each $x$ and $u$, in other words, $\Sigma$ is well-posed. For simplicity of notation, we often use $x(0) = x_0 (x_0 \in \mathcal{R}^n)$ as the initial state instead of the hybrid state $(I(0), x(0)) = (I_0, x_0) \in \{(I, x) \in \mathcal{I} \times \mathcal{R}^n \mid x \in S_I\}$, since by assumption, the value of the initial discrete state $I_0 \in \mathcal{I}$ is uniquely determined by each $x_0 \in \mathcal{R}^n$.

For $\Sigma$, the controllability notion is defined as follows.

Definition 1 For the system $\Sigma$, suppose that the final time $T \in \mathcal{N}_+$ and the direct product set $\mathcal{X} \subseteq \mathcal{R}^{2n}$ of the sets $X_0, X_T \subseteq \mathcal{R}^n$ of the continuous state ($\mathcal{X} := X_0 \times X_T$) are given. Then $\Sigma$ is said to be $(T, \mathcal{X})$-controllable if for each $x_0 \in X_0$, there exists an input vector sequence $U := \{u(0), u(1), \ldots, u(T-1)\} \in \mathcal{R}^{nT}$ satisfying $x(T) \in X_T$ under the initial state $x(0) = x_0$.

Now, we show a $(T, \mathcal{X})$-controllability condition for $T \in \{2, 3, \ldots\}$ (see [3, 4] for details of the deviation).  

\[1\] Different symbols have to be used for the cases $T = 1$ and $T \in \{2, 3, \ldots\}$; due to the limited space, the former is omitted.
Let $(I, X)$ be a sequence of the intermediate hybrid states $(I_k, x_k) \in \{(I, x) \in I \times \mathbb{R}^n | x \in S_I \}$, $(k = 1, 2, \ldots, T - 1)$, where $I := [I_1, I_2, \ldots, I_{T-1}]^T \in \mathbb{R}^{T-1}$ and $X := [x_1, x_2, \ldots, x_{T-1}]^T \in S_I$ for $S_I := S_{I_1} \times S_{I_2} \times \ldots \times S_{I_{T-1}}$ ($\subseteq$ is often called the discrete state trajectory). In addition, we define

$$E(I_0, x_0, I, X, x_T) := E^a_{i, I} \left[ \begin{array}{c} x_0 \\ X \\ x_T \end{array} \right] + E^b_{i, I},$$

where

$$E^a_{i, I} := \begin{bmatrix} E^a_{1, I} \\ E^a_{2, I} \\ \vdots \\ E^a_{T-1, I} \end{bmatrix}, \quad E^b_{i, I} := \begin{bmatrix} E^b_{1, I} \\ E^b_{2, I} \\ \vdots \\ E^b_{T-1, I} \end{bmatrix},$$

and $E^a_{i, I} := -(B^a_{i, I})^T A_{i, I}$, $E^b_{i, I} := (B^a_{i, I})^T$, $A_{i, I} := \text{rank}(B^a_{i, I}) B^a_{i, I} = n$ and $(B^a_{i, I})^T B^a_{i, I} = 0$ (an integer). Then we can prove that for any $T \in \{2, 3, \ldots\}$, $(I, x_0) \in \{(I, x) \in I \times \mathbb{R}^n | x \in S_I \}$, $I \subseteq \mathbb{R}^{T-1}$, and $\mathcal{X} \subseteq \mathbb{R}^n$, there exists a $U \subseteq \mathbb{R}^{mT}$ satisfying $[I(1) I(2) \cdots I(T-1)] = \mathbb{I}$ and $x(T) \in \mathcal{X}$ under the initial state $(I(0), x(0)) = (I_0, x_0)$ if and only if there exists an $[X^T X_T] \in S_I \times X_T$ satisfying

$$E(I_0, x_0, I, X, x_T) = 0.$$  

Thus the set of $x_0 \in \mathcal{X}$ for which there exists a $U \subseteq \mathbb{R}^{mT}$ satisfying $x(T) \in \mathcal{X}$ under the initial state $x(0) = x_0$ (i.e., $(I(0), x(0)) = (I_0, x_0)$) can be expressed as

$$\mathcal{X}_c(T, \mathcal{X}) := \bigcup_{I \in \mathbb{I}^{T-1}} \mathcal{X}^b_{c, I}(T, \mathcal{X}),$$

where

$$\mathcal{X}^b_{c, I}(T, \mathcal{X}) := \bigcup_{I \in \mathbb{I}^{T-1}} \mathcal{X}^b_{c, I}(T, \mathcal{X})$$

and

$$\mathcal{X}^b_{c, I}(T, \mathcal{X}) := \left\{ x_0 \in S_{I_0} \cap X_0 \mid \left\{ [X^T X_T]^T \in S_I \times X_T \mid E(I_0, x_0, I, X, x_T) = 0 \right\} \neq \emptyset \right\}.$$  

Note that $\mathcal{X}^b_{c, I}(T, \mathcal{X})$ expresses the set of $x_0 \in S_{I_0} \cap X_0$ for which there exists a $U \subseteq \mathbb{R}^{mT}$ satisfying $[I(1) I(2) \cdots I(T-1)] = \mathbb{I}$ and $x(T) \in \mathcal{X}$ under the initial state $(I(0), x(0)) = (I_0, x_0)$. Then we straightforwardly obtain a necessary and sufficient condition for $\Sigma$ to be $(T, \mathcal{X})$-controllable.

**Theorem 1** For $\mathcal{X}$, suppose that $T \in \{2, 3, \ldots\}$ and $\mathcal{X} \subseteq \mathbb{R}^{2n}$ are given. Then $\Sigma$ is $(T, \mathcal{X})$-controllable if and only if $\mathcal{X}_c(T, \mathcal{X}) = X_0$.

The condition in Theorem 1 can be checked by standard polyhedral manipulation techniques. However, such a deterministic approach will not be practical from the viewpoint of computational complexity; the required computation amount may grow exponentially with $T$ and $n$ [3]. So an alternative way for checking the $(T, \mathcal{X})$-controllability is required.

### 3. Probabilistic Controllability Analysis

In this section, the framework of the probabilistic controllability analysis, composed of the positive and negative one-sided error algorithms and the two-sided error algorithm, is introduced.

For simplicity of notation, we suppose that $T \in \{2, 3, \ldots\}$ and $\mathcal{X} \subseteq \mathbb{R}^{2n}$ are given in advance; so the symbols $\mathcal{X}^b_{c, I}$, $\mathcal{X}^b_{c, I}(T, \mathcal{X})$, are often used instead of $\mathcal{X}^b_{c, I}(T, \mathcal{X})$, $\mathcal{X}^b_{c, I}(T, \mathcal{X})$, respectively.

#### 3.1. Positive One-Sided Error Algorithm

The positive one-sided error algorithm [3], called Algorithm $P$, is given by the following procedure.

1. Given $T \in \{2, 3, \ldots\}$, $\mathcal{X} \subseteq \mathbb{R}^{2n}$, and $N_P \in N$.
2. $j := 1$; $\mathcal{X}_0(0) := \emptyset$.
3. while $j \leq N_p$ do
   2.1: Generate an i.i.d. random vector $v^j$ from a uniform distribution on $\mathbb{I}^{T-1}$.
   2.2: $\mathcal{X}_0(j) := \bigcup_{I \in \mathbb{I}^{T-1}} \mathcal{X}^b_{c, I}(T, \mathcal{X})$.
   2.3: $\mathcal{X}_0(j) := \mathcal{X}_0(j - 1) \cup \mathcal{X}_0(j)$.
   2.4: if $\mathcal{X}_0(j) = \mathcal{X}_0(j)$, then Halt: return "$Y$".
   3. end
   3. $j := j + 1$.
3. Halt: return "$Y$".

Algorithm $P$ approximately solves the $(T, \mathcal{X})$-controllability problem with a positive one-sided error as follows.

For given $x_0 \in \mathcal{X}_0$, consider the set of the discrete state trajectories defined by $\mathcal{I}^{T-1}(x_0) := \{ I \in \mathbb{I}^{T-1} | x_0 \in \mathcal{X}^b_{c, I}(T, \mathcal{X}) \}$. This expresses the set of $I \in \mathbb{I}^{T-1}$ for which there exists a $U \subseteq \mathbb{R}^{mT}$ satisfying $[I(1) I(2) \cdots I(T-1)] = \mathbb{I}$ as well as $x(T) \in \mathcal{X}$ under the initial state $x(0) = x_0$. So if $I$ is an i.i.d. random vector from a uniform distribution on $\mathbb{I}^{T-1}$, the relation

$$\operatorname{Prob}\{ I \in \mathbb{I}^{T-1}(x_0) \} = \frac{\operatorname{card}(\mathbb{I}^{T-1}(x_0))}{\operatorname{card}(\mathbb{I}^{T-1})}$$

holds. Then the following result is obtained.

**Lemma 1** For $\Sigma$, suppose that $T \in \{2, 3, \ldots\}$ and $\mathcal{X} \subseteq \mathbb{R}^{2n}$ are given. In addition, for Algorithm $P$, suppose that $N_P \in N$ is given by

$$N_p := \left[ \frac{\ln \frac{1}{\delta_P}}{\ln \frac{1}{1 - \delta_P}} \right]$$

for arbitrarily given $\delta_P, \delta_P \in (0, 1)$. Then if Algorithm $P$ outputs "$Y$", there then exists an $x_0 \in \mathcal{X}_0$ satisfying

$$\operatorname{Prob}\{ I \in \mathbb{I}^{T-1}(x_0) \} \leq \varepsilon_P \geq 1 - \delta_P.$$  

Otherwise, i.e., if Algorithm $P$ outputs "$Y$", $\Sigma$ is $(T, \mathcal{X})$-controllable.

Lemma 1 can be obtained from a well-known inequality in the probabilistic robustness analysis (see e.g., [6]). This implies that if Algorithm $P$ outputs "$N_\varepsilon$", there exists an $x_0 \in \mathcal{X}_0$ such that $\operatorname{card}(\mathbb{I}^{T-1}(x_0))/\operatorname{card}(\mathbb{I}^{T-1}) \leq \varepsilon_P$.
holds with probability more than or equal to $1 - \delta_P$; so if $\varepsilon_P$ and $\delta_P$ are sufficiently small, it is guaranteed with sufficiently high probability that for almost all $l \in \mathbb{T}^{-1}$, there does not exist a $U \in \mathbb{R}^{mT}$ satisfying $[I(l) I(2) \cdots I(T - 1)]^T = 1$ and $x(T) \in \mathcal{X}_T$ under the initial state $x(0) = x_0$. Note here that there does not exist a $U \in \mathbb{R}^{mT}$ satisfying $x(T) \in \mathcal{X}_T$ under the initial state $x(0) = x_0$ if and only if $\text{Prob}\{\exists l \in \mathbb{T}^{-1}(x_0)\} = 0$, and thus (10) can be considered as a kind of relaxed condition for $\Sigma$ with $x(0) = x_0$ to not have a $U \in \mathbb{R}^{mT}$ satisfying $x(T) \in \mathcal{X}_T$; hence if “$\Sigma$,” is outputted, it is approximately verified that $\Sigma$ is not $(T, \mathcal{X})$-controllable, more precisely, it is verified that $\Sigma$ is not $(T, \mathcal{X})$-controllable in the sense that there exists an $x_0 \in \mathcal{X}_0$ satisfying (10). On the other hand, if Algorithm P outputs “$Y^*$”, it is verified that $\Sigma$ is $(T, \mathcal{X})$-controllable.

### 3.2. Negative One-Sided Error Algorithm

Under assumption

(A1) The set $\mathcal{X}_0$ is bounded and measurable,

the negative one-sided error algorithm [4] (Algorithm N) is given as follows.

0: Given $T \in \{2, 3, \ldots\}$, $\mathcal{X} \subseteq \mathbb{R}^{2n}$, and $\mathcal{N}_P, \mathcal{N}_N \in \mathcal{N}$;
1: $i := 1$;
2: while $i \leq \mathcal{N}_N$
   2.1: Generate an i.i.d. random vector $x_0^i$ from a uniform distribution on $\mathcal{X}_0$;
   2.2: If $x_0^i \notin \mathcal{X}_0^i$, then Halt: return “$N^*$”;
   2.3: $i := i + 1$;
3: Halt: return “$Y^*$”;

In a similar way to the case of Algorithm P, it is shown that Algorithm N approximately solves the $(T, \mathcal{X})$-controllability problem with a negative one-sided error.

Let $x_0$ be a random vector from a uniform probability density function on $\mathcal{X}_0$, and we formally define

$$\text{Prob}\{x_0 \in \mathcal{X}_0\} := \frac{\text{vol}(\mathcal{X}_0)}{\text{vol}(\mathcal{X}_0)}.$$

Note that $\mathcal{X}_0 \subseteq \mathcal{X}_0$ and $\text{Prob}\{x_0 \in \mathcal{X}_0 - \mathcal{X}_0\} = 1 - \text{Prob}\{x_0 \in \mathcal{X}_0\}$. Then the following result is obtained.

**Lemma 2** For $\Sigma$, suppose that $T \in \{2, 3, \ldots\}$ and $\mathcal{X} \subseteq \mathbb{R}^{2n}$ satisfying (A1) are given. In addition, for Algorithm N, suppose that $\mathcal{N}_N \in \mathcal{N}$ is given by

$$\mathcal{N}_N := \left\lfloor \frac{\ln \frac{1}{\varepsilon_N}}{\ln \frac{1}{1 - \delta_N}} \right\rfloor$$

(12)

for arbitrarily given $\varepsilon_N, \delta_N \in (0, 1)$. Then if Algorithm N outputs “$N^*$”, then $\Sigma$ is not $(T, \mathcal{X})$-controllable. Otherwise, i.e., if Algorithm N outputs “$Y^*$”,

$$\text{Prob}\{\text{Prob}\{x_0 \in \mathcal{X}_0 - \mathcal{X}_0\} \leq \varepsilon_N\} \geq 1 - \delta_N$$

(13)

holds.

This implies that if Algorithm N outputs “$N^*$”, then it is verified that $\Sigma$ is not $(T, \mathcal{X})$-controllable. On the other hand, if Algorithm N outputs “$Y^*$,” with sufficiently small $\varepsilon_N$ and $\delta_N$, it is guaranteed with sufficiently high probability that for almost all $x_0 \in \mathcal{X}_0$, there exists a $U \in \mathbb{R}^{mT}$ satisfying $x(T) \in \mathcal{X}_T$ under the initial state $x(0) = x_0$. Thus it is approximately determined that $\Sigma$ is $(T, \mathcal{X})$-controllable, i.e., it is determined that $\Sigma$ is $(T, \mathcal{X})$-controllable in the sense of (13).

### 3.3. Two-Sided Error Algorithm

The two-sided error algorithm [5], called Algorithm T, is derived by mixing the positive and negative one-sided error algorithms in Sections 3.1 and 3.2, i.e., it is based on the random sampling of the discrete state trajectory $l \in \mathbb{T}^{-1}$ and the initial continuous state $x_0 \in \mathcal{X}_0$. This is shown as follows.

0: Given $T \in \{2, 3, \ldots\}$, $\mathcal{X} \subseteq \mathbb{R}^{2n}$, and $\mathcal{N}_P, \mathcal{N}_N \in \mathcal{N}$;
1: $i := 1$;
2: while $i \leq \mathcal{N}_N$
   2.1: Generate an i.i.d. random vector $x_0$ from a uniform distribution on $\mathcal{X}_0$;
   2.2: while $i \leq \mathcal{N}_P$
      2.2.1: $j := 1$;
      2.2.2: Generate an i.i.d. random vector $j$ from a uniform distribution on $\mathcal{T}^{-1}$;
      2.2.3: If $x_0^j \notin \mathcal{X}_0^j$, then go to 2.4;
      2.2.4: $j := j + 1$;
   2.3: Halt: return “$N^*$”;
   2.4: $i := i + 1$;
3: Halt: return “$Y^*$”;

Algorithm T repeatedly checks (in line 2.2 including lines 2.2.1–2.2.4) if or not an i.i.d. random vector $x_0^j \in \mathcal{X}_0$ satisfies the condition

(C1) $x_0^j \notin \mathcal{X}_0^j$ for all $j \in \{1, 2, \ldots, \mathcal{N}_P\}$.

If (C1) holds for some $i \in \{1, 2, \ldots, \mathcal{N}_N\}$, the algorithm outputs the symbol “$N^*$” otherwise, i.e., if (C1) does not hold for all $i \in \{1, 2, \ldots, \mathcal{N}_N\}$, it outputs the symbol “$Y^*$”. Then from Lemmas 1 and 2, the following result is straightforwardly obtained.

**Lemma 3** For $\Sigma$, suppose that $T \in \{2, 3, \ldots\}$ and $\mathcal{X} \subseteq \mathbb{R}^{2n}$ satisfying (A1) are given. In addition, for Algorithm T, suppose that $\mathcal{N}_P, \mathcal{N}_N \in \mathcal{N}$ are given by (9) and (12) for arbitrarily given $\varepsilon_P, \delta_P, \varepsilon_N, \delta_N \in (0, 1)$. Then if Algorithm T outputs “$N^*$”, then there exists an $x_0 \in \mathcal{X}_0$ satisfying (10). Otherwise, (13) holds.

Therefore, for given $\varepsilon_P, \delta_P, \varepsilon_N, \delta_N$ as accuracy parameters, Algorithm T under (A1) can determine with a two-sided error if $\Sigma$ is $(T, \mathcal{X})$-controllable or not.

Finally, Table 1 condenses the features and the computational complexities of the three algorithms (see [3, 4, 5] for the proofs of the computational complexities), where the computational complexities are estimated by the specification parameters of the $(T, \mathcal{X})$-controllability, i.e., $T$ and the
Table 1: Relation of the probabilistic algorithms.

<table>
<thead>
<tr>
<th>Random Variable in Algorithm</th>
<th>Algorithm P</th>
<th>Algorithm N</th>
<th>Algorithm T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solution of Algorithm</td>
<td>Positive Result</td>
<td>Deterministic sense</td>
<td>Probabilistic sense ((13) holds)</td>
</tr>
<tr>
<td></td>
<td>Negative Result</td>
<td>Probabilistic sense ($\exists x_0 \in X_0$ s.t. (10))</td>
<td>Probabilistic sense ($\exists x_0 \in X_0$ s.t. (10))</td>
</tr>
<tr>
<td></td>
<td>w.r.t. $T$</td>
<td>Polynomial</td>
<td>Exponential</td>
</tr>
<tr>
<td></td>
<td>w.r.t. $n$</td>
<td>Exponential</td>
<td>Polynomial</td>
</tr>
</tbody>
</table>

Table 2: Results ($T = 10$) by the three algorithms.

(a) Case $\zeta = 1$.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P</th>
<th>N</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(T, \mathcal{X})$-controllability</td>
<td>$N_\ast$</td>
<td>$N_\ast$</td>
<td>$N_\ast$</td>
</tr>
<tr>
<td>Computation time [sec]</td>
<td>122</td>
<td>1404</td>
<td>8</td>
</tr>
</tbody>
</table>

(b) Case $\zeta = 2$.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>P</th>
<th>N</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(T, \mathcal{X})$-controllability</td>
<td>$Y$</td>
<td>$Y$</td>
<td>$Y$</td>
</tr>
<tr>
<td>Computation time [sec]</td>
<td>3.9</td>
<td>124.7</td>
<td>136.1</td>
</tr>
</tbody>
</table>

Figure 1: Subregions $S_1$ in example.

on the computer with the Intel Pentium 4 2.53GHz processor and the 2048MB memory and the symbols “$Y$”, “N”, “$Y$′”, “N′” express the output of the algorithms. Note here that “$Y$′” means that $\Sigma$ is $(T, \mathcal{X})$-controllable in the sense that $\text{Prob}\{x_0 \in X_0 - X_0\} \leq 0.1 [\%]$ holds with probability more than or equal to 99.9 [\%], and “N′” does that $\Sigma$ is not $(T, \mathcal{X})$-controllable in the sense that there exists an $x_0 \in X_0$ satisfying $\text{Prob}\{1 \in I^{-1}(x_0)\} \leq 0.1 [\%]$ with probability more than or equal to 99.9 [\%]. The result is based on ten trials, where the algorithms answered the same result in every trial, and the computation time expresses the mean value in the trials (the variance was small).

This table shows that the $(T, \mathcal{X})$-controllability problem, which must not be solved in a deterministic way, can be solved within a practically short time by the proposed framework.

References


270
Controller Optimisation for Hybrid Systems by use of Return Maps

Jörg Krupar and Wolfgang Schwarz

Institute of Fundamentals of Electronics, Dresden University of Technology
Mommsenstrasse 13, 01062 Dresden, Germany
Email: krupar@iee.et.tu-dresden.de, schwarz@iee1.et.tu-dresden.de

Abstract—In this paper we present a method to optimise a controller for hybrid systems by use of a return map. We propose a criterion and an associated method based on numerical analysis that can be used to optimise parameters of a hybrid system controller. Using the presented method the controller parameters of a chaotically operating DC-DC boost converter in current programmed mode (CPM) are optimised.

1. Introduction

In the last decades hybrid systems became subject of interest in application as well as in research. Typical examples are DC-DC converters and antilock braking systems. In this paper we consider discretely controlled continuous systems. In these systems perpetually switching has to occur forever in order to assure correct function. Depending on the event generator the system can operate periodically or chaotically. The task of the controller is to maintain the continuous state within a certain interval.

Due to the possibility of complex system behaviour (e.g. bifurcations and chaos) the design of a reliable controller is sometimes a frustrating task. Classical methods for controller design in general require a locally stable control loop or at least a locally stable control loop model. As we want to include controller design for chaotically operating systems a statistical approach is reasonable. The method can be applied to controller design of periodically as well as chaotically operating hybrid systems. We use the return map model of [1] and apply statistical analysis to obtain the steady state distribution and the convergence to it. These two characteristics together are used in an optimisation criterion for controller design.

The paper is organised as follows. Section 2 summarizes the hybrid system model used in the paper and clarifies the design task. In section 3 we propose the optimisation criterion and associated controller optimisation method. Section 4 demonstrates the proposed method to optimise the controller parameters of a chaotically operating boost converter. Section 5 draws conclusions.

2. Hybrid system

Here we completely use the modeling approach of [1]. In this section we only repeat the most important issues.

The hybrid system is composed of a hybrid plant and a continuous controller \( R \), fig. 1. The hybrid plant consists of a continuous subsystem \( x = f(x, q) \), a discrete subsystem

\[ q(k + 1) = g(q(k), e) \]

and an event generator \( \Phi \). \( \Phi \) generates the switching events \( e \) and their occurrence times \( t \) from \( x, t, q \) and a control input \( u \) that modifies the parameters of switching event generation. Each time when an event occurred a state transition of the discrete subsystem is initialised.

The task of the continuous controller is to

- adjust the steady state \( \bar{x} \), characteristics such as the ensemble mean and variance according to application specific requirements
- move the continuous state \( x \) to \( \bar{x} \)
- hold the continuous state \( x \) within \( \bar{x} \)

Therefore \( R \) derives the control input \( u \) from the continuous state \( x \).

If the structure of \( R \) is given then for the complete system plant-controller the return map \( P \) can be derived.

\[
P : \bar{s}(m + 1) = P(\bar{s}(m)) \text{ with } \\
\bar{s}(m) = (\bar{x}(m), \bar{t}(m), \bar{q}(m))^T
\] (1)

determines the successor state \( \bar{s}(m + 1) \) from the actual one. A detailed description of the derivation and representation of the return map can be found in [1].

Throughout the paper we will use the return map to model this system. We completely reduce the investigation to the states at the switching events. As the resulting system is event discrete and the sequence of the states at certain switching events is obtained by repeated iteration of the...
return map we will call the resulting model an iterated system. The controller task in this model reduces to the control of $x$ at the observed switching events.

In the next section we propose a criterion to optimise controller parameters for a fixed controller structure.

3. A Method to Optimise Hybrid System Controller Parameters

3.1. Optimisation Criterion

Here we use a criterion that rates the controller tasks named in section 2. As here we want to include chaotically operating systems measure-based criteria are rather appropriated than trajectory-based criteria. A criterion that fulfills these requirements is

$$J = f_s \left( \mathbb{E}(\|\xi^2_t\|) \right) + \sum_{m=1}^{\infty} m^n \left( \mathbb{E}(\|\xi(m)\|^2) - \mathbb{E}(\|\xi_t^2\|) \right),$$

where $\xi(m) = \mathbb{x}(m) - \mathbb{x}_{ref}$ is the process error and $\mathbb{x}_r(m) = \mathbb{x}_r(m) - \mathbb{x}_{ref}$ is the steady state error.

The function $f_s$ rates the steady state error. To rate convergence to the steady state the deviation between the trajectory ensemble and the steady state is accumulated by the sum term. The power $p$ can be used either to assure exponential convergence of the criterion ($p < 1$) or, if exponential convergence is assured, to define to which extend the deviation from steady state with increasing number of iterations is penalised ($p > 1$).

We note this criterion is related to the well-known Bolza form

$$J(x_0, u) = x(t_e)'Sx(t_e) + \int_0^{t_e} (x(t)'Qx(t) + u(t)'Ru(t)) \, dt.$$  \tag{3}$$

3.2. Algorithm

The optimisation algorithm is similar to that used in [2]. An initial parameter set is used to obtain the return map. To limit the calculation effort a limited number of initial settings is used. Then by repeated iteration of the return map $P$ an ensemble of trajectories is obtained from which the expectation value in eq. (2) for each iteration $m$ is calculated. Each parameter is optimised sequentially using a newton-like algorithm to obtain the minimum of eq. (2). The algorithm stops if a minimum is found. The result is an optimal parameter setup for the controller.

4. Application Example

In this section the proposed optimisation method will be applied to a PI controlled boost converter in CPM. Fig. 2 depicts the converter. The corresponding functions of the hybrid plant model, fig. 1, are given in [1]. The PI controller $u = k_p \cdot \Delta V + \int_0^t \Delta V(t) \, dt$ with $\Delta V = V_{ref} - v_o(t)$ has a saturation of the integrator output voltage. The saturation values are $i_{min}$ and $i_{max}$. The converter parameters are $V_i = 4 \, \text{V}$, $L = 100 \, \mu \text{H}$, $C = 10 \, \mu \text{F}$, and $T_{clk} = 10 \, \mu \text{s}$, $V_{ref} = 15 \, \text{V}$. The peak inductor current is limited to $I_{max} = 5 \, \text{A}$. As a result of the input-to-output voltage ratio the converter operates chaotically [3].

To rate the results of the optimisation we first derive controller parameters by use of the linearised standard averaging model and phase margin criterion. The resulting controller is used as a reference. As the DC-DC converter operates chaotically the averaging method in its original form and linearising would result in a pole in the right half plane [4].

In this case no controller design for the chaotic system is possible. To overcome this problem we slightly modify the averaged model by setting $<i_L >_{\tau} = i_{ref}$. Then averaging and linearising results in the transfer function

$$\frac{V_o}{I_{ref}} = R V_{in} - sL \frac{V_o}{V_o(sCR + 2)}.$$  $R$

The controller parameters are adjusted to achieve about 60° phase margin and maximise crossover frequency. This leads to the parameter setup $k_p = 0.8 \frac{V}{\mu A}$, $k_i = 3000 \frac{A}{V}$. The integrator output saturation $i_{max}$, $i_{min}$ values are chosen as $i_{max} = 3 \, \text{A}$, $i_{min} = 0$.

To rate the resulting transient behaviour we use the density $f_{\nu_e}(\nu, m)$ of the output voltage. The initial settings are as follows: $10 \, \text{V} \leq v_o(0) \leq 20 \, \text{V}$, $0 \leq i_L(0) \leq i_{max}$, $i_{min} \leq i(t) \leq i_{max}$, where $i$ is the integrator output value of the controller.

The output voltage densities are depicted in fig. 3. De-
Figure 4: Mean value and standard deviation of the converter output voltage for phase margin controller design

Figure 6: Mean value and standard deviation of the converter output voltage for optimised parameters

Figure 5: Output voltage pdf for optimised controller parameters

Figure 7: Output voltage pdf for optimised controller parameters, 2nd parameter setup

Figure 8: Mean value and standard deviation of the converter output voltage for optimised parameters, 2nd parameter setup

Depending on the initial state the output voltage raises up to 26 V. The steady state is achieved after about 70 cycles.

Now the following optimisation cases are considered.

**Case 1)** $k_I$, $k_P$ and $I_{\max}$ are optimised. $f_2(x) = 500 \cdot x$ and $p = 1$, $i_{\min} = 0$ and $i_{\max} = 3 \text{ A}$ are kept constant.

This results in the parameter setup $I_{\max} = 2.41 \text{ A}$, $k_P = 0.45 \frac{\text{A}}{\text{V}}$, $k_I = 3715 \frac{\text{A}}{\text{V}}$. The corresponding output voltage densities are depicted in fig. 5. The steady state is achieved after about 40 cycles.

**Case 2)** $k_I$, $k_P$ $I_{\max}$ and $i_{\min}$ are optimised. $f_2(x) = 500 \cdot x$ and $p = 1$. The resulting parameter setup is $I_{\max} = 2.49 \text{ A}$, $k_P = 0.54 \frac{\text{A}}{\text{V}}$, $k_I = 7627 \frac{\text{A}}{\text{V}}$, $i_{\min} = i_{\max} = 0.9 \text{ A}$. $i_{\min} = i_{\max}$ means the integrator is not active but adds a constant value to the controller output. The resulting output voltage densities are depicted in fig. 7. The steady state is achieved after about 20 cycles. The steady state mean value error is 0.2 V.

**Case 3)** $f_3$ is increased into $f_3 = 50000 \cdot \epsilon$ to reduce steady state error. The resulting parameter setup is $I_{\max} = 2.5 \text{ A}$, $k_P = 0.21 \frac{\text{A}}{\text{V}}$, $k_I = 2390 \frac{\text{A}}{\text{V}}$, $i_{\min} = i_{\max} = 1.2 \text{ A}$. The output voltage densities are depicted in fig. 9. Here the
steady state mean value error is reduced to 0.02 V while the number of cycles to achieve steady state remained constant. Furthermore the variance of the steady state error is reduced from 0.1 V² to 0.03 V². Again the integrator saturation values result in a constant value output of the integrator.

All figures shown previously depict an offset between \( V_{\text{ref}} \) and the stationary state. The reason for this offset is that we used the event values in the figures and to optimise the controller parameters. As the state \( \mathbf{x} \) has a ripple, the event value is different from the average value. As the integral part of the controller removes the average value deviation the event values will always have an offset.

If the average or the event value is to be considered depends on the application. A partially consideration of the average can be done by adding an output function to the map \( P_0 \) to calculate the one-cycle averages from the event values. Then eq. (2) has to use the averages instead of the event values. If we use this criterion and run the optimisation procedure for fixed \( i_{\text{min}} = 0 \) and \( i_{\text{max}} = 3 \) A we obtain the parameter setup \( I_{\text{max}} = 1.6 \) A, \( k_P = 1.52 \frac{\text{A}}{\text{V}} \), \( k_I = 10672 \frac{\text{A}}{\text{Vs}} \). Fig. 11 depicts the corresponding mean value and standard deviation behaviour of the one-cycle averages. We note a small offset of the one-cycle averages remains. The reason for this behaviour is that here the one-cycle averages of the ensemble are equally weighted. A further improvement is possible by the use of time-interval weighted averages.

5. Conclusions

We proposed a measure based optimisation criterion and an optimisation method for controllers of hybrid systems by use of return maps. Nonlinear system behaviour such as saturation automatically is included in the return map and hence considered by the optimisation procedure. Furthermore in contrast to existing approaches this method can be applied to chaotically operating systems as shown in the CPM DC-DC converter example.

References


Supervisory Control of an Experimental Switched DES

M. Canu†, J. Haurogne†, D. Morel†, and N. Rakoto-Ravalontsalama†

†Department of Control, Ecole des Mines de Nantes
4, rue Alfred Kastler, F-44307 Nantes Cedex 03, France
Email: {mcunu, haurongne, morel, rakoto}@emn.fr

Abstract— The notion of switched discrete event systems (DES) has been introduced in [14]. This is a class of DES where each automaton is the composition of two basic automata, but with different composition operators. A switching occurs when there is a change of the composition operator, but keeping the same two basic automata. A mode behavior is defined as the active DES behavior for a given composition operator. Composition operators are supposed to change more than once so that each mode is visited more than once. In this paper we study the behavior of an experimental manufacturing as an example of switched DES.

1. Introduction

Supervisory control initiated by Ramadge and Wonham [15] provides a systematic approach for the control of discrete event system (DES) plant. Most of the properties of a given composed system depend on the composition operator. The modular approach reflects the underlying physical properties of complex systems such as manufacturing systems.

The most common composition operators used in supervisory control theory are the product and the parallel composition [1], [20]. However many different types of composition operators have been defined, e.g., the prioritized synchronous composition [3], the biased synchronous composition [7], see [19] for a review of most of the composition operators. Multi-Agent composition operator [17], [18] is another kind of operator, which differs from the synchronous product in the aspects of simultaneity and synchronization.

Related work concerns a) fault diagnosis for DES (the readers are referred to [4] for a comprehensive survey), b) mode-automata for reactive system programming introduced by Maraninchi and Remond [9], c) supervisory uniqueness for operating mode systems studied by Kamach et al. [6] where the authors propose a multi-model approach to DES, and finally d) sensor failure tolerant supervisory control proposed by Rohloff [16] where different automata are used to model the system observation behavior in the various modes of operations.

This paper studies the application of switched DES methodology to an Experimental Manufacturing Cell. This cell is composed of two robotized workstations connected to a central conveyor belt. Recently, three new semi-automated workstations have been added in order to increase the flexibility aspects of the cell. This flexibility allows the designer to study different mode behaviors of the experimental cell.

The paper is organized as follows. In Section 2, the notation and preliminaries are given. The notion of switched DES is recalled in Section 3. In Section 4, the controllability property is studied. Finally the experimental cell behavior is described in Section 5.

2. Notation and Preliminaries

Let the discrete event system plant be modeled by a finite state automaton [5]

$$G = (Q, \Sigma, \delta, q_0, Q_m)$$

where $Q$ is the finite set of states, $\Sigma$ is the finite set of events associated with the transitions in $G$, $\delta : Q \times \Sigma \rightarrow Q$ is the partial transition function, $q_0$ is the initial state and $Q_m \subseteq Q$ is the set of marked states.

Let $\Sigma'$ be the set of all finite strings of elements in $\Sigma$ including the empty string $\epsilon$. The function $\delta$ can be generalized to $\delta : \Sigma' \times Q \rightarrow Q$. The notation $\delta(s, q)$! for any $s \in \Sigma'$ and $q \in Q$ denotes that $\delta(s, q)$ is defined. Let $L(G) \subseteq \Sigma'$ be the language generated by $G$, that is,

$$L(G) = \{ s \in \Sigma' | \delta(s, q_0) \}$$

Let $K \subseteq \Sigma'$ be a language. The set of all prefixes of strings in $K$ is denoted by $K$ with $K = \{ s \in \Sigma' | \exists s' \in \Sigma^*; st \in K \}$. A language $K$ is said to be prefix closed if $K = K$. The event set $\Sigma$ is decomposed into two subsets $\Sigma_c$ and $\Sigma_u$ of controllable and uncontrollable events, respectively, where $\Sigma_c \cap \Sigma_u = \emptyset$. A controller, called a supervisor, controls the plant by dynamically disabling some of the controllable events. A closed language $K \subseteq L(G)$ is said to be controllable with respect to $L(G)$ and $\Sigma_u$ if [15]

$$K \cup L(G) \subseteq K$$

In the supervisory control theory, composition means synchronization of finite state automata. The basis for the definition of all the composition operators are $G_a = (Q_a, \Sigma_a, \delta_a, q_{0a}, Q_{ma})$ and $G_b = (Q_b, \Sigma_b, \delta_b, q_{0b}, Q_{mb})$ with disjoint state sets $Q_a \cap Q_b = \emptyset$ but generally overlapping event sets. The result of any composition is an automaton $G = G_a \parallel G_b = (Q, \Sigma, \delta, q_0, Q_m)$ with the state $Q = Q_a \times Q_b$, the event set $\Sigma = \Sigma_a \cup \Sigma_b$ and initial state $x_0 = x_{0a}, x_{0b}$, where $\parallel$ is a composition operator. Each operator is defined by a distinct transition function with $\sigma \in \Sigma$ a single event and $q \in Q$ a state.
Among the different types of composition operators, we recall here the biased synchronous composition (BSC) and the strict product composition (SPC).

**Definition 1** [7] The Biased Synchronous Composition (BSC) is defined as follows. The automaton $G_a$ is called the master and $G_b$ is called the follower.

$$
\delta(q, \sigma) = \begin{cases} 
\delta_a(q_a, \sigma) \times \delta_b(q_b, \sigma) & \text{if } \delta_a(q_a, \sigma)! \land \delta_b(q_b, \sigma)!
\delta_a(q_a, \sigma) \times \{q_b\} & \text{if } \delta_a(q_a, \sigma)! \land \neg \delta_b(q_b, \sigma)!
\emptyset & \text{otherwise.}
\end{cases}
$$

**Definition 2** The strict product composition (SPC) is defined as follows.

$$
\delta(q, \sigma) = \begin{cases} 
\delta_a(q_a, \sigma) \times \delta_b(q_b, \sigma) & \text{if } \delta_a(q_a, \sigma)! \land \delta_b(q_b, \sigma)!
\emptyset & \text{otherwise.}
\end{cases}
$$

These two composition operators will be taken as example in the next sections.

### 3. Switched DES

The basic idea is the following. Without loss of generality we consider two automata $G_a$ and $G_b$ as defined above. Let $G_i$ be the composed automaton from $G_a$ and $G_b$ with operator $\text{op}_i$, that is $G_i = G_a \text{|| op}_i G_b$. In the same way let $G_j$ be the composed automaton from the same $G_a$ and $G_b$ but with operator $\text{op}_j$, that is $G_j = G_a \text{|| op}_j G_b$, as it is depicted in Figure 2 and Figure 3.

**Definition 3** Equivalent states. The states $(q_a, q_b)$ of automaton $G_i$ and $(q_c, q_d)$ of automaton $G_j$ are said to be equivalent $(q_c, q_d)$ if they result from the composition of the same pair of states but with different composition operators $(q_a = q_c$ and $q_b = q_d)$.

**Assumptions.** Given two automata $G_i$ and $G_j$, switching between automaton $G_i$ and automaton $G_j$ is possible if the following assumptions hold.

1. $G_i$ and $G_j$ have at least two equivalent states
2. Switching between $G_i$ and $G_j$ is performed through their equivalent states.
3. Switching from $G_i$ to $G_j$ has zero duration, as well as from $G_j$ to $G_i$.

**Definition 4** [14] Switched DES. A switched discrete event system is defined as follows.

$$L_{\text{switched}}(G) = L(G_i), i \in I = \{1, \ldots, n\}$$

where $G_i$ is the model of DES $\nu$, and $I$ is an index set. In this special case, $G_i = G_a \text{|| op}_i G_b$.

We can see in Figure 2 and Figure 3 the automaton $G_1$ and $G_2$, respectively. Then automata $G_1$ and $G_2$ can switch between them, as it is shown in Figure 4. Actually, the switching are made through the equivalent states of $G_1$ and $G_2$, see Figure 5.

![Figure 1: Switched DES](image1)

![Figure 2: Automaton $G_1$](image2)

![Figure 3: Automaton $G_2$](image3)

![Figure 4: Switched automaton](image4)

![Figure 5: Automata switching through equivalent states](image5)

We give here below some examples of switched DES:

- Manufacturing systems where the operating modes are changing (e.g. from normal mode to degenerated mode)
- Discrete event systems after an emergency signal (from normal to safety mode)
- Complex systems changing from normal mode to recovery mode (or from safety mode to normal mode).
We can distinguish, like for the switched continuous-time systems, the notion of autonomous switching where no external action is performed and the notion of controlled switching, where the switching is forced. The notion of switched DES has been adapted from the switched continuous-time systems. For a survey of switched continuous-time systems, one can refer to [8] and the references therein.

On one hand DES and continuous-time systems share the notion of controllability (but each domain has its own definition). On the other hand, stability analysis in continuous-time systems cannot be adapted to DES (even though some work exist on the stability of DES [10], [11] and the references therein). Thus the notion of stability analysis has been changed to nonblocking analysis. Before defining the problems, we need to define the notion of switching sequence. A switching sequence is defined to be the successive active automata when the successive switchings occur. The following problems have been defined in [14]

- **Problem A.** Find conditions that guarantee that the switched DES (1) is controllable with respect to the Language \( L(G) \) and with respect to all the uncontrollable events, for any switching sequence.

- **Problem B.** Identify the classes of switching sequences for which the switched DES (1) is controllable with respect to the Language \( L(G) \) and with respect to all the uncontrollable events.

- **Problem C.** Find conditions that guarantee that the switched DES (1) is nonblocking.

- **Problem D.** Identify the classes of switching sequences for which the switched DES (1) is nonblocking.

We can note that discretization of a switched continuous system (see e.g., [12]) may be a solution to the adaptation to the DES context.

4. **Controllability of Switched DES**

In this section we address a specific problem related to the controllability of a switched DES (1).

**Problem 1.** Given a switched automaton \( L_{\text{switched}} = L(G) = L(G_i), i \in I = \{1, 2\} \) where \( G_1 = G_{1||G} G_1 = G_{1||GSC} G_2 \) and \( G_2 = G_{2||G} G_2 = G_{2||GSC} G_2 \), find the conditions that guarantee the controllability of the switched DES \( L_{\text{switched}} = L(G) \).

**Theorem 1** Given a switched automaton \( L_{\text{switched}} = L(G) = L(G_i), i \in I = \{1, 2\} \) where \( G_1 = G_{1||G} G_1 = G_{1||GSC} G_2 \) and \( G_2 = G_{2||G} G_2 = G_{2||GSC} G_2 \), the switched automaton \( L(G) = L(G_i) \) is controllable with respect to both \( L(G_1) \) and \( L(G_2) \) and with respect to \( \Sigma c \) if

1. \( K_B, K_a, L(G_2) \) are pairwise non conflicting
2. \( K_B, L(G_a) \) are non conflicting

3. \( K_B \) is controllable w.r.t. \( L(G_a) \)

**Proof.** The proof can be found in [14]. It is based on four propositions that have been given in [19].

5. **Experimental Manufacturing Cell**

An automated manufacturing system generally consists of a number of interconnected material processing stations capable of processing a wide variety of part types, a material transport system, a communication system for integrating all aspects of manufacturing and a supervisory control system. The experimental manufacturing cell is composed of the following components [2]: a) a central conveyor belt, b) two robotized workstations, with a station conveyor each, c) a transfer system between the central conveyor belt and the station conveyor, d) another transfer system between the station conveyor and the corresponding robot, and e) a load-unload robotized workstation.

Recently, three semi-automated workstations have been added to increase the flexibility aspects of the cell. Indeed, each semi-automated workstation can perform either manual or automated tasks. The experimental manufacturing cell is depicted in Figure 6.

Behavioral specifications of such an automated manufacturing system include: a) logic-based specifications (e.g. safety, error recovery, the sequencing of operations, part routing and production volume requirement), b) temporal production specification: production times, and c) utility optimality specification: e.g. costs.

6. **Conclusions**

This paper studies the application of the switched DES methodology, introduced in [14] to an Experimental Manufacturing Cell. The different mode behaviors were possible to obtain thanks to the recently added semi-automated workstations. These latter increased the flexibility of the system, and it allows the designer to apply the switched DES approach. Future work will be focused on a) obtaining more different mode behaviors for controllability, and b) study nonblocking properties in some specific cases.

7. **Acknowledgement**

This work is supported in part by the French 2000-2006 "Contrat Etat-Région CER STIC 9, N. 18036: Optimisation des processus industriels” Nantes, France.

**References**


Figure 6: Experimental Manufacturing Cell


Control-invariance of hybrid systems with sampled-data state feedback control

Yoshiyuki Tsuchie† and Toshimitsu Ushio†

†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531 JAPAN

Abstract—In this paper, we consider a hybrid system controlled by a sampled-data state feedback controller whose action is periodically time-driven, that is, the control inputs can change only at particular points in time of a synchronized global time base. We introduce transition systems as semantics of the controlled hybrid systems. Moreover, we consider a control problem given by a predicate and show that there always exists the suprema control-invariant subpredicate for any predicate.

1. Introduction

A hybrid automaton is widely used as a model of hybrid systems[1]. A computer-controlled system is an example of hybrid systems since it has both continuous and discrete variables associated with the physical process (the plant) and the logical dynamics (the control logic and external environment), respectively[2]. In actual systems, the digital controller is time-driven, that is, the measurements and subsequent control actions are periodically executed and the control input value is constant in intersamplings. If the sampling frequency is fast relative to the physical process, it may not be necessary to model the sampling process. However, as the sampling period approaches time constant inherent in the process dynamics, the sampling period has much effect on the system.

In discrete event systems, a state feedback controller is often used as a logical control problem, where the control specification is given by a predicate on their states[3]. Its control action is determined by their current states. A discrete event system is called control-invariant if there exists a state feedback controller such that all reachable states in the closed-loop system controlled by the controller satisfy the predicate. A necessary and sufficient condition for the system to be control-invariant is derived[3]. The state feedback control for the discrete event systems is extended to hybrid systems[4] and hybrid automata with forcible events[5]. Forcible events are events that can be forced to occur by the control so that temporal performance can be improved[6]. Ushio and Takai extend transition semantics of uncontrolled hybrid systems[1] to that of controlled hybrid systems with forcible events and show necessary and sufficient conditions for a predicate to be control-invariant.

In this paper, we consider a hybrid automaton with a sampled-data state feedback controller which updates its control action periodically.

The rest of this paper is organized as follows: Section 2 reviews transition systems, several predicate transformations, and a concept of control-invariance. Section 3 introduces a hybrid automaton with the sampled-data state feedback controller and two transition systems as its semantics. Section 4 introduces two transition systems to describe the feasible behaviors of a controlled hybrid automaton. Finally, Section 5 discusses the control-invariance of the hybrid automaton with the sampled-data state feedback controller and shows that there exists the suprema control-invariant subpredicate.

2. Preliminaries

We use a labeled transition system $T=(Q, Act, \mathcal{T}, Q_0)$ in order to define semantics of hybrid systems, where $Q$ is a set of states, $Act$ is a set of labels, $\mathcal{T}\subseteq Q \times Act \times Q$ is a state transition relation, $Q_0 \subseteq Q$ is a set of initial states. $Act(T; q) = \{a \in Act | \exists q' \in Q \text{ s.t. } (q, a, q') \in \mathcal{T}\}$. Let $\mathcal{P}(Q)$ be the set of all predicates on $Q$. A predicate $P$ is true at state $q \in Q$ if $P(q)=1$, and false if $P(q)=0$. Denoted by $\vee$, $\wedge$, and $\neg$ are disjunction, conjunction, and negation of predicates, respectively. The term “predicate” and “subset”($=\{q \in Q | P(q)=1\}$) can be used interchangeably. A partial order “$\leq$” for $\mathcal{P}(Q)$ is defined as follows: for $P_1, P_2 \in \mathcal{P}(Q)$, $P_1 \leq P_2 \iff P_2(q) \leq P(q)$ for each $q \in Q$. For each $a \in Act$, a predicate $D_a$ is defined by

$$D_a(q) = \begin{cases} 1 & \text{ if } a \in Act(T; q), \\ 0 & \text{ otherwise).} \end{cases} \tag{1}$$

We define predicate transformations $wp_a: \mathcal{P}(Q) \rightarrow \mathcal{P}(Q)$ and $wlpa: \mathcal{P}(Q) \rightarrow \mathcal{P}(Q)$ as follows:

$$wp_a(p) = \begin{cases} 1 & \text{ if } Post(q, a) \neq 0 \text{ and } (\forall q' \in Post(q, a)) P(q') = 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

$$wlpa = \vee_{a \in Act} wp_a \tag{3}$$

where $Post(q, a) = \{q' \in Q | (q, a, q') \in \mathcal{T}\}$. For a subset $A \subseteq Act$, we define $wp_a(P) = \vee_{a \in A} wp_a(P)$. For a subset $A \subseteq Act$, $P \in \mathcal{P}(Q)$ is said to be $(T; A)$-invariant iff, for any $a \in A$, $P \leq wlpa$.}

3. Hybrid automaton and Sampled-data state feedback controller

We consider a hybrid system $H$ modeled by a hybrid automaton with forcible events which can be forced to occur
by external control actions as follows:

\[ H = (V, E, \Sigma, \text{event}, \text{init}, \text{Flow}, \text{jump}), \] (4)

where

- \( V \) is the set of nodes;
- \( \Sigma \) is the set of events and partitioned into the sets \( \Sigma_c, \Sigma_u \) of controllable and uncontrollable events, respectively. Moreover, let \( \Sigma_r \) be the set of forcible events. For simplicity, we assume that any forcible event is controllable;
- \( E \subseteq V \times \Sigma \times V \) is the set of edges with associated events, that is, \( e(v, \sigma, v') \) is an edge \( e \in E \) from a node \( v \) to a node \( v' \) labeled by event \( \sigma \) and corresponds to a discrete transition by the occurrence of \( \sigma \);
- For each node \( v \in V \), \( \text{inv}(v) \subseteq \mathbb{R}^n \) is the set of values which the continuous variable can take in \( v \), and \( \text{flow}(v) \subseteq \mathbb{R}^n \times \mathbb{R}^n \) is a set of values which \((x, \dot{x})\) can take in \( v \).
- \( \text{init} \) assigns the initial continuous states, that is, \( \text{init}(v) \) is the set of all possible initial continuous states in node \( v \); and
- For each edge \( e(v, \sigma, v') \), \( \text{jump}(e) \subseteq \mathbb{R}^n \) is the jump relation, that is, \( (x, x') \in \text{jump}(e) \) means that the continuous state \( x \in \text{inv}(v) \) jumps to \( x' \in \text{inv}(v') \) when \( \sigma \) occurs.

Note that \( \Sigma_c \cap \Sigma_u = \emptyset \) and \( \Sigma = \Sigma_c \cup \Sigma_u \). The state set \( Q_H \) of a hybrid automaton \( H \) is given by \( Q_H = \{(v, x) | v \in V, x \in \text{inv}(v)\} \). Let \( \text{guard}(e) \) be an occurrence condition of the discrete transition by edge \( e(v, \sigma, v') \) \( \in E \):

\[
\text{guard}(e) = [x \in \text{inv}(v)] \Rightarrow [x', x'] = \text{jump}(e)].
\]

Let \( \mathcal{F}(P, \delta, v, x, x') \) be a set of functions \( F : [0, \delta] \rightarrow \mathbb{R}^n \) satisfying the following conditions:

1. \( F(0) = x \) and \( F(\delta) = x' \);
2. For every \( e \in (0, \delta) \), \( F(e) \in \text{inv}(v) \) and \( (F(e), F(\delta)) \in \text{flow}(v) \);
3. For any \( \varepsilon_i, \varepsilon_j \in (0, \delta) \), \( P(v, F(e_i)) = P(v, F(e_j)) \).

Moreover, \( \mathcal{F}(P, \delta, v, x, x') = \bigcup_{e \in \text{inv}(v)} \mathcal{F}(P, \delta, v, x, x') \).

**Assumption 1** We assume that, for any \( e(v, \sigma, v') \in E \) and \( \sigma \in \Sigma_r \), \( \text{guard}(e) \) is a closed set.

In a conventional sampled-data control system, sampling times are periodic and each control input value is periodically changed. In this paper, let sampling times be periodical time sequence with a period \( T \) and an initial phase 0.

Thus, we introduce a sampled-data state feedback controller \( f \) as follows:

\[ f = (f_1, f_2), \] (5)

where

- \( f_1 : Q_H \rightarrow \Gamma_1 \) gives a set of control-enabled events, where \( \Gamma_1 = \{ \gamma | \Sigma_c \subseteq \gamma \subseteq \Sigma \} \); and
- \( f_2 : Q_H \rightarrow \Gamma_2 \) is a set of forcible events which are control-enabled and forced to occur, where \( \Gamma_2 = 2^{\Sigma_r} \).

Denoted by \( H^f \) is a hybrid automaton \( H \) with forcible events controlled by \( f \).

Henzinger introduces two transition systems, called timed and time-abstract transition systems, in order to represent semantics of \( H[1] \). Uschio and Takai extended them to a controlled hybrid automaton with a control specification given by a predicate [5]. Next, we extend them to a hybrid automaton with a sampled-data state feedback controller.

To define semantics of this automaton by transition system, the transition system must have state variables which indicate a control pattern assigned by the controller at the latest sampling time, duration between the current time and the latest sampling time, and the current state in \( Q_H \).

Thus, we can define two transition systems as semantics of a hybrid automaton with a sampled-data state feedback controller \( f \).

(1) A sampled-data controlled timed transition system

\[ \mathcal{H}(H^f, P) = (Q^f, \text{Act}_H, T^f, Q^f_0), \] (6)

where \( Q^f_0 \subseteq Q_H \times \Gamma_1 \times \Gamma_2 \times [0, T] \) is the set of states in the transition system and \( Q^f_0 = \{ (q_0, \gamma_1, \gamma_2, t_0) | q_0 \in Q_H, f(q_0) = (\gamma_1, \gamma_2) \} \) is the initial state set. Each element of a state \( (q_1, \gamma_1, \gamma_2, t_0) \in Q^f_0 \) is defined as follows: \( q_1 \) indicates a state of \( H^f \), \( \gamma_1 \) is an elapsed time from the latest sampling time, \( \gamma_2 \in \Gamma_1 \) and \( \gamma_2 \in \Gamma_2 \) are control patterns assigned at the latest sampling time by the controller \( f \).

\( \text{Act}_H \subseteq \Sigma \cup [0, T] \) is the set of events. \( T^f \subseteq Q_H \times \text{Act}_H \times Q^f_0 \) is defined as follows: Consider \( q_1 = (q, \gamma_1, \gamma_2, t_0) \), \( q_1 = (q', \gamma_1', \gamma_2', t') \in Q^f_0 \), where \( q = (v, x) \) and \( q' = (v', x') \).

\((A)\) For \( \sigma \in \text{Act}_H \cap \Sigma \), \( (q_1, \sigma, q_2) \in T^f \) if the following conditions are satisfied:

1. \( 3e(v, \sigma, v') \in E \) s.t. \( (x, x') \in \text{jump}(e) \), and
2. \( \sigma \in \gamma_1, \gamma_1 = \gamma_1', \gamma_2 = \gamma_2', \) and \( \omega = \omega' \).

\((B)\) For \( \delta \in [0, T] \), \( (q_1, \delta, q_2') \in T^f \) if the following conditions are satisfied:

1. \( v = v' \) and \( x, x' \in \text{inv}(v) \),
2. if \( \delta = \omega' - \omega \), then there exists \( F \in \mathcal{F}(P, \delta, v, x, x') \) such that the following conditions are satisfied:

   \((a)\) \( F(\delta) = x', \gamma_1 = \gamma_1', \) and \( \gamma_2 = \gamma_2' \),
   \((b)\) for \( t \in [0, \delta) \), \( F(e) \in E \), if \( F(t) \in \text{guard}(e) \), then \( \sigma \notin \gamma_2 \).
3. if \( \delta = \omega' - \omega \), then all the following conditions are satisfied:

   \((a)\) \( \delta = 0, \omega = T, \) and \( \omega' = 0, \)
   \((b)\) \( q = q' \) and \( f(q') = (\gamma_1', \gamma_2') \).

(2) A sampled-data controlled time-abstract transition system

\[ \mathcal{H}(H^f, P) = (Q^f, \text{Act}_H, T^f, Q^f_0), \] (7)

where the state sets \( Q^f_1 \) and \( Q^f_0 \) are same as those of the sampled-data controlled timed transition system. \( \text{Act}_H \subseteq \Sigma \cup \{ \tau_a, \tau_c \} \), where \( \tau_a, \tau_c \notin \Sigma \) are events. \( T^f \subseteq Q^f_0 \times \mathbb{R}^n \).
Actsa × Qf is defined as follows: transitions related to events in Σ are same as those in T f in H f. Consider q1=(q, γ1, γ2, ω), q′ s = (q′, γ′1, γ′2, ω′) ∈ Qf. Let Δ(q1, q′ s) = ∅ ∈ [0, T] (q1, δ, q′ s) ∈ T f.

(A) (q1, τ, q′ s) ∈ T f if Δ(q1, q′ s)≠∅, ω = T, and ω′ = 0.

(B) (q1, τ, q′ s) ∈ T f if Δ(q1, q′ s)≠∅ and 0 < ω < ω′.

4. A sampled-data transition system with feasible inputs

In this section, we define two transition systems to describe the feasible behavior of a controlled hybrid automaton H.

(1) A sampled-data timed transition system with feasible inputs

\[ \mathcal{F}(P) = (Qs, Actsa, Tsa, Qsa), \] (8)

where Qs = Q0 × Γ1 × Γ2 × [0, T] is the set of states in the transition system and Q0 = [(q0, γ1, γ2, 0) ∈ Q] is the initial state set. Each element of a state (q, γ1, γ2, ω) ∈ Qs is defined as follows: q indicates a state of H. ω is an elapsed time from the latest sampling time. γ1 ∈ Γ1 and γ2 ∈ Γ2 are control patterns hold at the latest sampling time.

Actsa ⊆ Σ ∩ [0, T] is the set of events. Tsa is defined as follows: Consider q1 = (q, γ1, γ2, ω), q′ s = (q′, γ′1, γ′2, ω′) ∈ Qs, where q = (ν, x) and q′ = (ν′, x′) ∈ Qs ∩ Σ.

(A) for σ ∈ Actsa ∩ Σ, (q1, σ, q′ s) ∈ Tsa if the following conditions are satisfied:

1. \( 3e(v, σ, v′) ∈ E \) s.t. (x, x′) ∈ jump(e), and

2. \( σ ∈ γ1, γ1 = γ′1, γ2 = γ′2, \) and \( ω = ω′. \)

(B) for δ ∈ [0, T], (q1, δ, q′ s) ∈ Tsa if the following conditions are satisfied:

1. \( ν = ν′ \) and \( x, x′ \) ∈ inv(ν).

2. if \( δ = ω′ − ω \), then there exists F ∈ \( \mathcal{F}(P, δ, v, x) \) such that the following conditions are satisfied:

(a) \( F(δ) = x′, γ1 = γ′1, γ2 = γ′2. \)

(b) for \( t ∈ [0, δ] \) and \( E intervenes (δ, t, v) ∈ E, \) if \( F(t) \) is guard(e), then \( δ ≠ γ2. \)

3. if \( δ ≠ ω′ − ω \), then the following condition is satisfied:

(a) \( \delta = 0, q = q′, ω = T, \) and \( ω′ = 0. \)

(2) A sampled-data time-abstract transition system with feasible inputs

\[ \mathcal{F}sa(P) = (Qs, Actsa, Tsa, Qsa), \] (9)

where the state sets Qs and Qsa are same as those of the sampled-data controlled timed transition system. Actsa ⊆ Σ ∪ τs, τc. Consider q1 = (q, γ1, γ2, ω), q′ s = (q′, γ′1, γ′2, ω′) ∈ Qs. Let Δ(q1, q′ s) = ∅ ∈ [0, T] (q1, δ, q′ s) ∈ Tsa.

(A) (q1, τ, q′ s) ∈ Tsa if Δ(q1, q′ s)≠∅, ω = T, and ω′ = 0.

(B) (q1, τ, q′ s) ∈ Tsa if Δ(q1, q′ s)≠∅ and 0 < ω < ω′.

We extend a predicate P on QH to Qsa as follows: for P : Qs → [0, 1] is defined by P(qs) = P(q) for each state q1 = (q, γ1, γ2, ω) ∈ Qs.

From the above definitions, the following lemmas are easily proved.

Lemma 1 Let f be a sampled-data state feedback controller for H. Then, for any qs, q′ s ∈ Qf, and event a ∈ Σ ∪ [τs, τc],

\[ (q1, a, q′ s) ∈ Tsa \] in \( \mathcal{F}(H f(P)) \)

\[ ⇔ (q1, a, q′ s) ∈ Tsa in \mathcal{F}(P), \] (10)

\[ (q1, τ, q′ s) \] in \( \mathcal{F}(H f(P)) \)

\[ ⇒ (q1, τ, q′ s) ∈ Tsa in \mathcal{F}(P). \] (11)

Lemma 2 In \( \mathcal{F}sa(H f(P)) \) for any qs ∈ Qs,

\[ |Post(qs, τ)| ≤ 1 in \mathcal{F}sa(H f(P)), \] (12)

where |A| is the number of elements in the set A.

5. Supremal control-invariant subpredicate

A concept of control-invariance plays an important role in state feedback control of discrete event systems[3]. A predicate Pe \( \mathcal{P}(Qs) \) is said to be control-invariant if there exists a sampled-data state feedback controller f such that P is \( \mathcal{F}(H f(P), Actsa) \)-invariant. Such a controller f is called a permissive feedback controller. In general, a given predicate P ∈ \( \mathcal{P}(Qs) \) is not necessarily control-invariant. In this section, we propose a procedure for computation of the supremal control-invariant subpredicate. We introduce some definitions for the predicate P as follows:

- \( C I(P) \) is the set of all control-invariant subpredicates of \( P ∈ \mathcal{P}(Qs) \).
- \( 0 ∈ \mathcal{P}(Qs) \) is the predicate such that, for any qs ∈ Qs, \( 0(qs) = 0 \). Since \( 0 ∈ C I(P), C I(P) ≠ ∅. \)
- A predicate \( P^1 ∈ C I(P) \) called the supremal control-invariant subpredicate of P is defined as follows: for any \( P^1 ∈ C I(Qs), P ≤ P^1. \)

Ushio and Takai showed that there always exists \( P^1 \) for the hybrid systems with forcible events[5]. In this section, we show the same property also holds for the hybrid automaton with the sampled-data state feedback controller.

Theorem 1 Let I be any index set. If \( P_i ∈ \mathcal{P}(Qs) \) is control-invariant for each i ∈ I, then, \( P_I = \bigvee_{i ∈ I} P_i \) is control-invariant.

By Theorem 1, there exists its supremal control-invariant subpredicate \( P^1 \) for any predicate \( P ∈ \mathcal{P}(Qs) \).

We define swp and swlp as follows: For a given labeled transition system T, its state qs, and predicate P,

\[ swp(P)(qs) = \begin{cases} \ 1 & Post(qs, a) ≠ ∅ in T and \ 3e(qs) ∉ Post(qs, a) \ s.t. P(qs) = 1, \ (13) \\
\ 0 & otherwise, \end{cases} \]

\[ swlp(P) = ~Da ∨ swp(P) \] (14)
In addition, we define a controller set $F$ as follows: for a predicate $P \in \mathcal{P}(Q_H)$,
\[
F(P) = \{ f : Q_H \rightarrow \Gamma_1 \times \Gamma_2 | \text{for each } q \in Q_H, \quad f(q) = (\gamma_1, \gamma_2) \text{ s.t. } P(q, \gamma_1, \gamma_2, 0) = 1 \} \tag{15}
\]

The following theorem gives an iterative scheme for computing the supremal control-invariant subpredicate.

**Theorem 2** For any $P \in \mathcal{P}(Q_H)$, consider the following iterative computation: $P_{l+1} = P_l \wedge \Psi(P_l)$ $(l \geq 0)$, where $P_0 := P$. Then the following implication holds:

If there exists $k \geq 0$ such that $P_{k+1} = P_k$, then $P^* = P_k$ and each $f^* \in F(P_k)$ is a permissive controller, where $\Psi : \mathcal{P}(Q) \rightarrow \mathcal{P}(Q)$ is defined as follows:
\[
\Psi(P) = \bigwedge_{\sigma \in \tau_r, \tau_s} wlp_{\sigma}(P) \wedge swlp_{\tau}(P) \in \mathcal{P}(P). \tag{16}
\]

**Proof:** Assume there exists $k$ such that $P_{k+1} = P_k$. For the above iterative scheme, we have the following conditions: In $\mathcal{P}(P_l)$, for any $a \in \Sigma \cup \{\tau_r\}$,
\[
P_{k+1} = P_k = P_k \wedge \Psi(P_k) \leq wlp_{\sigma}(P_k). \tag{17}
\]
\[
P_{k+1} = P_k = P_k \wedge \Psi(P_k) \leq swlp_{\tau}(P_k). \tag{18}
\]

Then, from Lemma 1 and the definition of $F(P_k)$, the following conditions hold: for any $f^* \in F(P_k)$,
\[
P_k \leq wlp_{\sigma}(P_k) \in \mathcal{P}(Q), \quad wlp_{\sigma}(P_k) \leq wlp_{\sigma}(P_k) \in \mathcal{P}(Q(H), P_k). \tag{19}
\]
\[
P_k \leq swlp_{\tau}(P_k) \in \mathcal{P}(Q), \quad swlp_{\tau}(P_k) \leq swlp_{\tau}(P_k) \in \mathcal{P}(Q(H), P_k). \tag{20}
\]

Thus, $P_k$ is shown to be $(\mathcal{P}(Q(H), P_k) : \Sigma \cup \{\tau_r, \tau_s\})$-invariant. So $f^*$ is the permissive controller and $P_k$ is a control-invariant subpredicate i.e. $P_k \in \mathcal{P}(P)$, which implies $P_k \leq P^*$.

Next, we prove that $P^1 \leq P_l$ for $l = 0, 1, \ldots, k$ by induction. (1) $l = 0$. Since $P^1$ is a subpredicate of $P$, we have $P^1 \leq P_0$. (2) Suppose that $P^1 \leq P_l$ holds. Then, we show by a contradiction that $P^1 \leq P_{l+1}$ holds. If $P^1 \leq P_{l+1}$ does not hold, then there exists $q_0 \in Q$, such that $P^1(q_0) = 1$ and $\Psi(P_l)(q_0) = 0$ since $P^1(q_0) = 1 \Rightarrow P_l(q_0) = 1$ for any $q_0 \in Q$. If $\Psi(P_l)(q_0) = 1$, it holds, one of the following cases always holds: (a) In $\mathcal{P}(P_l)$, there exists $\sigma \in \Sigma$ such that $wlp_{\sigma}(P_l)(q_0) = 0$. Since $D_\sigma$ in $\mathcal{P}(P_l)$ is equivalent to that in $\mathcal{P}(P)$, there exists $q'_0 \in Q$, such that $(q_0, \sigma, q'_0) \in D_\sigma$ in $\mathcal{P}(P)$ and $P_l(q'_0) = 0$, which implies that $P^1(q'_0) = 0$. This contradicts the assumption that $P^1$ is control-invariant. (b) $wlp_{\sigma}(P_l)(q_0) = 0$ holds in $\mathcal{P}(P)$). Then there exists $q'_0 \in Q$, such that $(q_0, \tau_u, q'_0) \in \mathcal{D}^\omega_{\tau_u}$ in $\mathcal{P}(P)$ and $P_l(q'_0) = 0$. Since $P^1$ is control-invariant and $P^1(q_0) = 1$, we have $(q_0, \tau_u, q'_0) \in \mathcal{D}^\omega_{\tau_u}$ in $\mathcal{P}(P^1)$. Since $P^1(q'_0) = 0$, this contradicts the assumption that $P^1$ is control-invariant. (c) In $\mathcal{P}(P)$, $swlp_{\tau}(P)(q_0) = 0$. Since $D_{\tau} = 1$ in $\mathcal{P}(P)$, it is shown that, for any controller $g$, $wlp_{\tau}(P^1)(q_0) = 0$ in $\mathcal{P}(Q(H), P^1)$. This contradicts the assumption that $P^1$ is control-invariant. For the above cases, we have $P^1 \leq P_{l+1}$. Therefore, we have $P^1 \leq P^*.$

Note that $P^1$ computed by the above scheme depends on time and a control pattern holds at the latest sampling time in general while the control specification $P \in \mathcal{P}(Q_H)$ is independent of time.

Practically, the iterative computation in Theorem 2 is implemented by using a bisimulation, and its termination is closely related to the existence of a finite bisimulations [7]. For hybrid automata without finite bisimulations, finite approximations of the bisimulations are needed to design a state feedback controller [8].

6. Conclusion

This paper considered the sampled-data state feedback control of a hybrid automaton with forcible events as a model of computer-controlled systems where control specifications are given by predicates.

We introduced two transition systems as semantics for the hybrid automata with the sampled-data state feedback controller and defined two transition systems with feasible inputs, showed that there always existed the supremal control-invariant subpredicate for any predicate.

In general, sampling times in practical control systems are fluctuated by computational delay or data transmission delay. So, it is future work to model this fluctuation as jitter in sampling times and discuss control-invariance of this model.

**References**


Consistency of a laser system driven by chaos and noise signals

Atsushi Uchida†, ‡, Shigeru Yoshimori†, Ryan McAllister‡ and Rajarshi Roy‡

†Department of Electronics and Computer Systems, Takushoku University
815-1, Tatemachi, Hachioji-shi, Tokyo 193-0985, Japan
‡ Institute for Research in Electronics and Applied Physics, University of Maryland
College Park, Maryland 20742, USA
Email: a-uchida@es.takushoku-u.ac.jp

Abstract – We experimentally and numerically investigate consistency of response output in a microchip laser driven repeatedly by chaos and colored noise with different bandwidths. An increase of consistency is observed as the amplitude of the drive signal is increased for both chaos and colored noise drive signals. After reaching an optimal value the consistency begins to decrease due to the deterministic dynamics of the nonlinear response as shown by the change in sign of the conditional Lyapunov exponent.

1. Introduction

Many nonlinear dynamical systems have an ability to generate consistent outputs when driven by a repeated external signal. “Consistency” is defined as the reproducibility of response waveforms in a nonlinear dynamical system driven repeatedly by a signal, starting from different initial conditions of the system. Consistency of dynamics is essential for information transmission in biological and physiological systems and for reproduction of spatiotemporal patterns in nature. Consistency tests could be applied in non-invasive diagnostic procedures to detect changes in system parameters due to aging, catastrophic events or other system changes.

Several phenomena related to consistency have been studied in various nonlinear dynamical systems. Generalized synchronization has been observed [1-3] in which there is a functional relation between the dynamics of a drive and response system, but the dynamics may differ greatly in character. If one now couples two independent but identical response systems with the drive system under generalized synchronization, the response systems, starting from different initial conditions, display identical synchronization after transients have disappeared. Noise-induced synchronization is a phenomenon where two identical nonlinear systems driven by a common noise signal can be identically synchronized to each other [4]. Reliability of spike timing in neurons has been investigated [5], where neurons that are repeatedly driven by a random drive signal can fire a consistent spike train with high temporal precision. The characteristics of these three examples of drive-response systems are contained within the more general concept of consistency. Reproducibility of the response outputs with respect to a repeated drive signal is essential for all three phenomena.

In stochastic resonance [6] and coherence resonance [7], it is the periodicity or coherence of the system response that is of interest, and this is modified by an external noise signal. Here, the focus is on the reproducibility or consistency of the system response to a repeated, complex waveform drive signal. The response signal may or may not have a functional relationship to the drive signal; its consistency is a measure of the ability of the external drive to interact with and excite the system degrees of freedom in a reproducible fashion. If the drive signal is too weak, it cannot overcome the effect of internal noise sources; if it is too strong, it may deterministically destabilize the response of the nonlinearly coupled degrees of freedom.

The concept of consistency is illustrated in Fig. 1. Any complex waveform such as deterministic chaos or stochastic noise can be used as a drive signal. This drive signal is sent repeatedly to a nonlinear dynamical system (called the response system) starting from arbitrary initial conditions. Complex temporal waveforms of the response system are obtained at each repetition of the drive signal. Consistency can be defined as the ability of a system to produce identical response outputs after some transient period, when the system is driven by a repeated drive signal. In this Letter we introduce a quantitative measure of consistency, and experimentally and numerically demonstrate its measurement in the dynamics of a physical laser system driven repeatedly by a complex waveform. We observe and explain three regimes of response - a growth in consistency as the drive signal amplitude increases, followed by optimal consistency and then a decrease of consistency as the drive amplitude increases still further. We believe that these are general
Fig. 2 Experimental results of temporal waveforms of (a) the chaotic drive signal and (b) the colored noise drive signal, and two corresponding response laser outputs. Numerical results of the temporal waveforms of (c) the chaotic drive signal and (d) the colored noise drive signal, and two corresponding response laser outputs. Consistent outputs are observed for all the figures after transient.

2. Experiment

2.1. Setup

To test consistency, we used chaos and noise waveforms as drive signals. Our nonlinear dynamical system was a laser-diode-pumped Neodymium-doped Yttrium Aluminum Garnet (Nd:YAG) microchip laser, similar to that used for the observation of generalized synchronization of chaos [3], except no feedback loop was used. Two longitudinal modes are observed in the output of the laser in a wide range of pump power. An acousto-optic modulator (AOM) is inserted in the laser cavity to modulate the loss of the laser cavity. The digitized drive signal stored in a computer is sent to the AOM in the laser system through an amplifier and a low-pass filter to smooth the signal by using an arbitrary function generator connected with the computer. The response laser system is driven repeatedly by the same drive signal and temporal waveform of the response laser output is detected by using a digital oscilloscope and photodiode. We compare the detected signals for different repetitions to observe consistency of the response waveforms.

2.2. Temporal waveforms

First we used a chaotic signal generated by the same laser system with closed feedback loop as a drive signal [3]. The chaotic signal is sent to the response laser repeatedly. Temporal waveforms of the chaotic drive signal and two response laser outputs obtained from our experiment are shown in Fig. 2(a). Two consistent response outputs are clearly observed after a transient of $\sim 1$ ms in Fig. 2(a), even though the drive and response signals are totally different. We next used a colored noise signal generated by a numerical algorithm [8] as a drive signal and sent the noise signal to the laser system repeatedly. The inverse of the correlation time for the colored noise is set to be 40 kHz. Temporal waveforms obtained from our experiments are shown in Fig. 2(b). Consistent outputs of the response laser driven by the same colored noise signal are obtained after transients as well.

Corresponding numerical results for chaotic and colored noise drive signals are shown in Figs. 2(c) and 2(d), respectively. Two response outputs starting from different initial conditions converge to the consistent outputs after transients in both cases. These numerical results agree well with our experimental observation shown in Figs. 2(a) and 2(b). We observed two longitudinal laser modes in our experiments, resulting in a dynamical system with eight degrees of freedom. The numerical calculations to explain our experimental observations and to investigate the detailed characteristics of consistency were performed on the Tang-Statz-deMars (TSD) model which includes the effect of spatial hole burning in a two-mode laser [3,9].

2.3. Consistency

To investigate the characteristics of consistency, we quantitatively define the consistency parameter $C$ as the cross-correlation of two temporal response waveforms normalized by the product of their standard deviations: i.e.,

$$C = \frac{\langle (I_A - \langle I_A \rangle)(I_B - \langle I_B \rangle) \rangle}{\sigma_A \sigma_B},$$  

where $I_A$ and $I_B$ are the total intensities of the two response waveforms, $\langle I_A \rangle$ and $\langle I_B \rangle$ are the mean values of the two response waveforms, and $\sigma_A$ and $\sigma_B$ are the standard deviations of the two response waveforms. The angle brackets denote time averaging.

We calculated the time evolution of the consistency $C$ from the time series shown in Figs. 2(a) and 2(b). The consistency is averaged over 0.1 ms and calculated continuously. The amplitude of the drive waveform is measured by its standard deviation, $\sigma$. The results are shown in Fig. 3. For the curves indicated by the label
"consistent", the consistency parameter gradually increases during the transient region over the first 1 ms and stays at a value close to 1 after that, for both chaos and colored noise drive waveforms. However, when the amplitude of the drive waveforms is increased, the consistency fluctuates between -1 and 1 as shown in the curves indicated by the label "inconsistent." The achievement of consistency is dependent on the amplitude of the drive signal.

We varied the amplitude of the drive waveform to measure the characteristics of consistency. We computed the average of the consistency parameter over 13 ms after the transient period. Our experimental results for the consistency as a function of the amplitude of the drive waveform is shown in Fig. 4(a) for chaos and colored noise drives. There is a maximum of the consistency curve for both drive signals as the amplitude is increased. At the very small amplitude region ($\sigma < 0.02$), consistent outputs are not observed because the internal noise-driven relaxation oscillations are dominant in the laser output. As the amplitude is increased, the drive signal overcomes the internal noise and optimal, consistent outputs are observed. When the amplitude of the drive signal is increased further, the consistency decreases and inconsistent outputs are observed. Our numerical results indicate a similar curve with the maximum peak of consistency as shown in Fig. 4(b) (solid curve) for the chaos drive signal. In order to explain the development of inconsistency in the case of strong driving, the maximal conditional Lyapunov exponent along the trajectory of the response laser output is also estimated without the internal noise terms (dotted curve of Fig. 4(b)) [1,3]. We found that inconsistency appears at large drive amplitudes when the sign of the conditional Lyapunov exponent changes from negative to positive. This result suggests that dynamical response of the nonlinear system generates inconsistency for large amplitude drive waveforms. On the other hand, inconsistency appears at small drive amplitudes even though the conditional Lyapunov exponent is negative. In this region, internal noise predominates and prevents consistency. Consistency is thus optimized between these two regimes for some intermediate amplitude of drive waveform.

We define the threshold for "inconsistency" as the drive amplitude at which the sign of the conditional Lyapunov exponent changes from negative to positive values. The threshold is dependent on the type of drive waveform. Our numerical results show that the threshold for the chaos drive signal ($\sigma_{th,chaos} = 0.061$) is smaller than that for the colored noise drive signal ($\sigma_{th, colored} = 0.20$) shown as seen in Figs. 4(b) and 4(c). Since the chaotic signal is generated from the same laser system, it excites the dynamics of the laser system more effectively than the colored noise signal. This chaotic excitation thus leads more quickly to instability and reduces the threshold for inconsistency.
with the chaotic drive signal than with the colored noise drive. In the limit of zero correlation time of the colored noise ($\tau_c \rightarrow 0$), which represents white noise, we do not find a threshold for inconsistency as shown in Fig. 4(c). This point is further clarified when we change the correlation time $\tau_c$ for the colored noise systematically. The threshold for inconsistency for the colored noise is plotted as a function of $1/\tau_c$ as shown in Fig. 4(d) (solid curve). The threshold value changes as $1/\tau_c$ is changed. The curve of the threshold for inconsistency displays a clear minimum around a frequency of 80 kHz which corresponds to the relaxation oscillation frequency of the laser. As $1/\tau_c$ differs further from the relaxation oscillation frequency, the inconsistency threshold value becomes larger. In the high bandwidth (white noise) limit, the inconsistency threshold appears to become infinite and consistency is always maintained with strong driving.

We also measure the amplitude of the response laser outputs at the threshold for inconsistency in Fig. 4(d) (dotted line). It is worth noting that the amplitude of the response outputs is almost constant when $1/\tau_c$ of the colored noise is changed as shown in Fig. 4(d). Therefore the threshold for inconsistency is dominated by the amplitude of the response laser output, $\sigma = 0.18$, i.e., by the nature of its deterministic dynamics in response to the drive signal.

3. Conclusion

We have quantitatively defined a measure of consistency for the response of nonlinear systems to external drive waveforms. We have experimentally and numerically determined the consistency of response output in a laser driven repeatedly by different complex waveforms - chaos and colored noise with different bandwidths. An increase of consistency is observed as the amplitude of the drive signal is increased for both chaos and colored noise drive signals. After reaching an optimal value which is dependent on the internal noise strength, the consistency begins to decrease due to the deterministic dynamics of the nonlinear response as shown by the change in sign of the conditional Lyapunov exponent. These aspects of consistency of response to drive waveforms may be general features that can be observed in many driven nonlinear classical and quantum systems.

References


Consistency in time-delay systems with periodic feedback functions

Kazuyuki Yoshimura, Jun Muramatsu, and Peter Davis

Abstract—It is often observed that a dynamical system reproduces a consistent output in response to a repeatedly applied external input signal. This phenomenon is called consistency. We study the consistency in one-dimensional time-delay systems. It is analytically shown that the consistency occurs in a class of time-delay systems, which have continuous and periodic feedback functions. Numerical evidence for this analytical result is also presented.

1. Introduction

In a variety of dynamical systems, it is often observed that a system reproduces a consistent output in response to a repeatedly applied external input signal: i.e., the output waveform does not depend on the initial condition but is determined only from the external signal waveform. We call this phenomenon the identical consistency (IC). Various types of signals are possible as the external input. It depends not only on the system but also on the type of external input signal whether the IC occurs in the system. In the case of noise input, experimental evidence for the IC has been found for several systems as diverse as lasers [1], neuronal networks [2], and ecological systems [3]. On the other hand, there is also analytical and numerical evidence. It was analytically shown that the IC can occur in a wide class of limit cycle oscillators [4]. The IC was numerically demonstrated in chaotic maps [5], in chaotic differential equations [6], and in one-dimensional time-delay systems [7].

The above works have well established that the IC is a quite general phenomenon observed in a variety of dynamical systems. In the real world, due to feedback loops or memory effects, time delay is ubiquitous in nature and technology. Various chaotic optical systems have time delays due to their optical or electrical feedback loops [8, 9, 10]. The IC in those systems may be of interest from the point of view of information theoretical security. The method known as secret key agreement by public discussion [11] has been developed to generate secret keys from correlated randomness. It may be possible to apply this method to the correlated randomness generated in separate chaotic optical devices subjected to a common noise input signal. Therefore, the study of the IC in time-delay systems is of high importance. However, most of previous works consider non-time-delay systems and the IC in time-delay systems has not been well understood although some numerical results are presented in [7]. One of the fundamental problems is to clarify the class of time-delay systems that exhibit the IC. However, there has been still no result on this fundamental problem. In the present paper, we theoretically show that the IC occurs in a wide class of one-dimensional time-delay systems, which have continuous and periodic feedback functions. This theoretical result is confirmed by numerical experiments.

2. Time-delay systems and conditional Lyapunov exponent

We investigate one-dimensional time-delay systems with an external noise input of the form

$$\beta^{-1}\dot{x}(t) = -x(t) + \alpha f(x(t-1)) + \sqrt{2D} \xi(t),$$  \(1\)

where \(x \in \mathbb{R}\), \(f\) is a nonlinear function, \(\alpha, \beta, D\) are positive constants, and \(\xi\) represents normal Gaussian white noise. The constant \(D\) represents the noise intensity. The noise \(\xi\) has the properties \(\langle \xi(t) \rangle = 0\) and \(\langle \xi(s)\xi(s+t) \rangle = \delta(t)\), where \(\delta\) is Dirac’s delta function and \(\langle \cdot \rangle\) denotes averaging over the realizations of \(\xi\). The time delay is set as unity without loss of generality because this is always possible by rescaling the variables and the parameters. We place the following assumptions on the function \(f\): (i) \(f(x)\) is a continuous periodic function of \(x\) with the period \(L\), (ii) \(f(x)\) is piecewise smooth, i.e., \(f(x)\) is continuously differentiable in \([0, L]\) except a finite number of points, and (iii) there exist constants \(M_0\) and \(M_1\) such that \(|f(x)| \leq M_0\) and \(|f'(x)| \leq M_1\).

Consider two trajectories \(x(t)\) and \(y(t)\) of Eq. (1) with different initial conditions for the same realization of the noise \(\xi(t)\). The IC is said to occur if \(\lim_{t \to \infty} |x(t) - y(t)| = 0\) for any initial conditions. A precise criterion for the IC can be obtained in terms of the linear stability of the trajectory \(x(t)\): the IC occurs when \(x(t)\) is linearly stable.

Suppose that \(x(t)\) and \(y(t)\) have an infinitesimally small difference at the initial. Let \(\Delta(t)\) be defined by \(\Delta(t) = y(t) - x(t)\), which represents the small deviation. Since \(f\) is differentiable almost everywhere from the assumption (ii), the time evolution of \(\Delta\) is governed by the linearized equation obtained from Eq. (1):

$$\dot{\Delta}(t) = \beta \left[ -\Delta(t) + \alpha f'(x(t-1))\Delta(t-1) \right].$$  \(2\)

The presence of noise \(\xi\) affects the time evolution of the trajectory \(x\). Therefore, the influence of the noise appears...
through \(x\) in Eq. (2). A quantity called the largest conditional Lyapunov exponent (LCLE) \([12]\) is defined by
\[
\lambda = \lim_{t \to -\infty} \frac{1}{t} \ln \left(\frac{\int_{-t}^{0} \Delta^2(s) ds}{\int_{-t-1}^{0} \Delta^2(s) ds}\right)^{1/2}.
\] (3)
The trajectory \(x(t)\) is linearly stable if \(\lambda < 0\).

3. Theory

We apply the Krasovskii-Lyapunov functional approach \([13]\) to show that the IC occurs in the time-delay systems when \(\alpha\) is small. Equation (2) has the trivial solution \(\Delta(t) = 0\). The stability of this solution is equivalent to the linear stability of \(x(t)\). We introduce the non-negative-valued Krasovskii-Lyapunov functional
\[
V(t) = \Delta^2(t) + \beta \int_{t-1}^{t} \Delta^2(s) ds.
\] (4)
The Krasovskii-Lyapunov theory tells that the solution \(\Delta(t) = 0\) is stable if the derivative of \(V(t)\) along the solution of Eq. (2) is always negative. The derivative is given by
\[
\dot{V}(t) = \beta (-\Delta^2 + 2\alpha f'(x(t-1))\Delta x - \Delta^2),
\] (5)
where \(\Delta = \Delta(t)\) and \(\Delta_1 = \Delta(t-1)\). The right hand side of Eq. (5) is a quadratic form of the two variables \(\Delta\) and \(\Delta_1\) and it is always negative except for the case \(\Delta = \Delta_1 = 0\) if \((\alpha f')^2 < 1\), where we used \(\beta > 0\). Using \(\alpha > 0\), we have the condition \(\alpha |f'| < 1\). Since \(|f'|\) is bounded from the assumption (iii), the inequality \(\alpha |f'| < 1\) holds for any \(t \geq 0\) if \(\alpha\) satisfies the condition
\[
\alpha < \sup_{x \in \mathbb{R}} |f'(x)| = \alpha_c.
\] (6)
This gives a sufficient condition for the stability of the solution \(\Delta(t) = 0\). For arbitrary noise intensity \(D\), the IC occurs in time-delay systems defined by Eq. (1) if \(\alpha\) is small and the inequality (6) holds. We note that the condition (6) cannot guarantee the IC for a chaotic time-delay system because if \(\alpha\) is in the region given by Eq. (6), any solution \(x(t)\) of Eq. (1) is stable for \(D = 0\), that is, the noise-free system is not chaotic.

In what follows, we show that the IC occurs for any given \(\alpha (\geq \alpha_c)\) when \(e^{-\beta} \ll 1\) and the noise intensity \(D\) is sufficiently large. This includes the case of chaotic noise-free systems. If we introduce the new variable \(X = x/\sqrt{2D}\), then equation (1) is rewritten as
\[
\dot{X}(t) = -X(t) + \frac{\alpha}{\sqrt{2D}} f(\sqrt{2D}X(t-1)) + \xi(t).
\] (7)
Since \(f(\sqrt{2D}X(t-1))\) in the right hand side is bounded due to the assumption (iii), the second term can be neglected for large \(D\). If we neglect the corresponding term in the right hand side of Eq. (1), we obtain
\[
\dot{x} = -\beta x(t) + \beta \sqrt{2D} \xi(t).
\] (8)
This equation indicates that \(x(t)\) is approximated by the Ornstein-Uhlenbeck (OU) process for large \(D\). Then, the time correlation function is given by
\[
\langle x(x + t) \rangle_x = D \beta \exp[-\beta |t|].
\] (9)
The steady probability distribution of \(x\) is given by
\[
P_0(x) = \frac{1}{\sqrt{2\pi DB}} \exp\left[-\frac{x^2}{2DB}\right].
\] (10)
For any fixed \(t\), the event \(\xi(x) \leq t \leq x + dx\) occurs with the probability \(P_0(x)dx\).

Consider a generic solution \(x(t)\) of Eq. (1) and a nonzero solution \(\Delta(t)\) of Eq. (2). We measure the stability of the solution \(\Delta(t) = 0\) via the time evolution of \(V\) along these solutions. Since \(0 \leq \Delta^2 \leq V\) from Eq. (4), exponential decrease of \(V\) implies that \(\Delta(t)\) converges to zero and thus the solution \(\Delta(t) = 0\) is stable.

Let \(r\) and \(\theta\) \((0 \leq \theta < 2\pi)\) be new variables defined by \(\Delta(t) = r(t) \cos \theta\) and \(\Delta(t-1) = r(t) \sin \theta\). Then, equation (5) reads
\[
\dot{V}(t) = \beta (-1 + \alpha f'(x(t-1)) \sin 2\theta) r(t)^2.
\] (11)
Let \(\rho\) and \(\psi\) be defined by \(\rho(t) = r(t)^2 / V(t)\) and \(V(t) = \exp(\psi(t))\). We note that both of \(r^2\) and \(V\) exhibit either exponential increase or decrease but \(\rho\) exhibits neither of them since \(r^2\) and \(V\) are of the almost same order and \(\rho\) is defined by their ratio. Equation (11) is rewritten as follows:
\[
\dot{\psi}(t) = \beta (-1 + \alpha f'(x(t-1)) \sin 2\theta) \rho(t).
\] (12)
The exponential growth rate \(\lambda_V\) of \(V\) is defined by \(\lambda_V = \lim_{t \to -\infty} (\langle \psi(T) - \psi(0) \rangle / T)\). We assume the existence of \(\lambda_V\). Integrating Eq. (12), we have
\[
\lambda_V = \lim_{t \to -\infty} \frac{\beta}{T} \int_0^T \left[-1 + \alpha f'(x(t-1)) \sin 2\theta\right] \rho(t) \, dt.
\] (13)
We assume that this time average can be replaced by the ensemble average over the realizations of \(\xi\). Consider the ensemble of the solutions \(x\) and \(\Delta\) for all possible realizations of \(\xi\). Let \(P(x, \theta, \rho)\) be the steady joint probability distribution such that for any given \(t\) the events \(x \leq x(t-1) \leq x + dx, \theta \leq \theta(t) \leq \theta + d\theta,\) and \(\rho \leq \rho(t) \leq \rho + d\rho\) occur simultaneously with the probability \(P(x, \theta, \rho) dx d\theta d\rho\). Note that the argument of \(x\) is \(t-1\). Using the ensemble average, we have
\[
\lambda_V = \lim_{t \to -\infty} \frac{\beta}{T} \int_0^T \langle \rho \rangle \, dt.
\] (14)
where \(\langle \cdot \rangle\) denotes averaging over the probability distribution \(P(x, \theta, \rho)\): i.e., the average \(\langle A \rangle\) is defined by \(\langle A \rangle = \int_0^T \int_{x(t-1)} dx d\theta d\rho \, A(x, \theta, \rho) P(x, \theta, \rho)\) for an arbitrary function \(A(x, \theta, \rho)\).

Let us consider the time evolution of \(\Delta\) over an interval \([t-1, t]\), provided that \(x\) and \(\Delta\) are given over \([t-2, t-1]\). Solving Eq. (2), we can obtain
\[
\Delta(t) = \Delta(t-1) e^{-\beta + \alpha \int_{t-2}^{t-1} f'(x(s)) \Delta(s) e^{-\beta(s-1)} ds}.
\] (15)
The order of the second term can be estimated as $O(\Delta (t - 1)\mu/\alpha_c)$ if we evaluate as $f' = O(\Delta (t - 1)/\alpha_c)$. Since $e^{-\beta t} \ll 1 \leq \alpha/\alpha_c$, the first term can be neglected. In the second term, the dominant contribution to the integral comes from $s \in I$, where $I = [t - 1 - \tau_0, t - 1]$ with $\tau_0 = O(\beta^{-1})$, because of the factor $e^{-\beta(t-1)}$. Therefore, $\Delta(t)$ is approximately determined from the time series of $f'(x(s))\Delta(s)$ over the interval $I$.

Suppose that $D$ is large. In this case, the solution $x(s)$ is approximated by the OU process as shown in Eq. (8). Equation (9) shows that the correlation decay time can be estimated as $\tau_0 = O(\beta^{-1})$. Given $\Delta(t)$ for the initial interval $[-1, 0]$, $\Delta(s)$ for $s \in I$ is completely determined from the time series of $x(s)$ for $s \in [-1, t - 2]$. Since $\tau_0 < 1$ follows from $e^{-\beta t} \ll 1$, $x(s)$ for $s \in [-1, t - 2]$ is uncorrelated with $x(s')$ for $s' \geq t - 1$. Consequently, $\Delta(s)$ for $s \in I$ is uncorrelated with $x(s')$ for $s' \geq t - 1$.

Let $\phi(t)$ be the time correlation function between $f'(x(s))$ and $x(s+t)$ defined by

$$
\phi(t) = \frac{\langle f'(x(s))\rangle x(s+t)\varepsilon}{\sqrt{\langle f'(x(s))\rangle^2 \cdot \langle x(s) \rangle^2 \varepsilon^2}},
$$

where $\mu$ is the average of $f'(x)$. Let $\tau$ be the correlation decay time such that $\phi(t)$ is smaller than a certain small value $\varepsilon_\tau$ for $t > \tau$. For example, $\varepsilon_\tau = 0.01$. By definition of $\tau$, $f'(x(s))$ for $s \leq t - 1 - \tau$ is uncorrelated with $x(s')$ for $s' \geq t - 1$. Therefore, $f'(x(s))$ for $s \in I'$, where $I' = [t - 1 - \tau_0, t - 1 - \tau]$, has no correlation with $x(s')$ for $s' \geq t - 1$.

If $\tau \ll \tau_0$ holds, $I - I' = [t - 1 - \tau, t - 1]$ is a small fraction of $I$ and the contribution to the integral coming from $I - I'$ is negligible in Eq. (15). Therefore, we have $\Delta(t) \approx \alpha \beta \int_{I'} f'(x(s))\Delta(s)e^{-\beta(t-1)/\beta} ds$. Since both of $f'(x(s))$ and $\Delta(s)$ for $s \in I'$ are uncorrelated with $x(s')$ for $s' \geq t - 1$ as shown above, there is no correlation between $\Delta(t)$ and $x(s')$ for $s' \geq t - 1$.

The time $t$ is arbitrary in the above argument. Thus, it follows that $\Delta(s)$ for any $s \in [t - 1, t]$ is uncorrelated with $x(t - 1)$. Since by definition $\theta(t)$ and $\mu(t)$ are the variables determined from $\Delta(s)$ over $s \in [t - 1, t]$, they are also uncorrelated with $x(t - 1)$. Therefore, the stochastic variables $(\theta, \rho)$ are independent of $x$ in the joint probability distribution $P(x, \theta, \rho)$, i.e., we can approximate as $P(x, \theta, \rho) \approx P_1(x)P_2(\theta, \rho)$, where $P_1(x)$ asymptotically approaches $P_0(x)$ given by Eq. (10) as $D$ increases. If we use this approximation in Eq. (14), we can obtain

$$
\lambda = \beta \left[ -\langle \rho \rangle + \alpha \langle f'(x) \rangle \cdot \langle \rho \sin 2\theta \rangle \right],
$$

for large $D$. It follows from the definition of the average (i.e.) that $|\langle \rho \sin 2\theta \rangle| \leq |\langle \rho \sin \theta \rangle| \leq |\langle \rho \rangle| = \langle \rho \rangle$, where we used the fact $\rho \geq 0$ in the last equality. Using the inequality $|\langle \rho \sin 2\theta \rangle| \leq \langle \rho \rangle$ in Eq. (17), we arrive at

$$
\lambda \leq \beta \left[ -1 + \alpha \langle f'(x) \rangle \right] \langle \rho \rangle.
$$

If $\lim_{D \to \infty} \langle f'(x) \rangle = 0$ holds, then for any given $\alpha$ we have $-1 + \alpha \langle f'(x) \rangle < 0$ when $D$ is sufficiently large. In this case, the right-hand side in Eq.(18) is negative since both of $\beta$ and $\langle \rho \rangle$ are positive. This indicates that $\lambda = 0$ holds for large $D$. The negative $\lambda$ implies that the solution $\Delta(t) = 0$ is stable and thus the IC occurs.

We show that the two conditions $\tau \ll \tau_0$ and $\lim_{D \to \infty} \langle f'(x) \rangle = 0$ are satisfied in the relevant class of time-delay systems. As the function $f(x)$ is periodic, $f'(x)$ can be expanded by the Fourier series

$$
f'(x) = \sum_{m=1}^{\infty} a_m \sin \frac{2\pi m}{L} x + b_m \cos \frac{2\pi m}{L} x,
$$

where $a_m$ and $b_m$ are the Fourier coefficients. The time correlation function $\phi(t)$ can be obtained as follows:

$$
\phi(t) = \frac{\psi(D)}{\sigma(D)} \cdot \exp[-\gamma t],
$$

where $\sigma(D) = \sqrt{(\langle f'(x(s)) - \mu \rangle^2)/\varepsilon}$ and $\psi(D)$ is defined by

$$
\psi(D) = \sqrt{D\gamma} \sum_{m=1}^{\infty} \frac{2\pi m a_m}{L} \exp \left[ -\frac{2\pi^2 m^2}{L^2} D \right].
$$

Equation (20) shows that $\phi(t)$ decreases monotonically as $t$ increases since $\sigma$ and $\psi$ do not depend on $t$. For a given $\varepsilon_\tau$, the correlation decay time is obtained from Eq. (20) as

$$
\tau = \frac{1}{\gamma} \ln \frac{\psi(D)}{\varepsilon_\tau \sigma(D)}.
$$

It can be shown that $\sigma(D)$ converges to the finite value $(1/L) \int_0^L |f'(x)|^2 dx$ in the limit $D \to \infty$. As for $\psi(D)$, we have $\lim_{D \to \infty} \psi(D) = 0$ from Eq. (21). Therefore, the correlation decay time $\tau$ given by Eq. (22) becomes arbitrarily small as $D$ increases. This implies that the condition $\tau \ll \tau_0$ is satisfied for large $D$.

As for the average of $\langle f'(x) \rangle$, we have

$$
\langle f'(x) \rangle = \sum_{m=0}^{\infty} \int_{(m-1)L}^{mL} f'(x)P_0(x)dx,
$$

for large $D$ since $P_1(x)$ is approximated by $P_0(x)$. The inequality $|P_0'(x)| \leq 1/\sqrt{2\pi \beta D\rho}$ holds from Eq. (10). This indicates that $P_0(x)$ can be approximated by a constant over each interval $[(m-1)L, mL]$ for large $D$ in Eq. (23). If we use this approximation, we obtain

$$
\langle f'(x) \rangle = \frac{1}{L} \int_0^L f'(x) dx = \frac{1}{L} \left[ f(L) - f(0) \right] = 0,
$$

where we used the assumption (i): the continuity of $f(x)$ was used in the second equality and the periodicity in the third equality. Equation (24) shows $\lim_{D \to \infty} \langle f'(x) \rangle = 0$.

According to the preceding theory, $\lambda$ must have negative for large $D$ since the two conditions $\tau \ll \tau_0$ and $\lim_{D \to \infty} \langle f'(x) \rangle = 0$ are satisfied. Thus, it may be concluded that the IC occurs for any given $\alpha (\geq \alpha_c)$ in the class of time-delay systems satisfying (i)-(iii) if $e^{-\beta t} \ll 1$ and $D$ is sufficiently large.
4. Numerical results

We show some numerical results to demonstrate the above theoretical results. Two examples of the function \( f \) are employed: the first one is

\[
f(x) = \sin x,
\]

which is called the Ikeda system and describes some chaotic optical systems [8, 9], and the second one is

\[
f(x) = \begin{cases} 
  x - 2n & \text{if } 2n \leq x < 2n + 1, \\
  2(n+1) - x & \text{if } 2n + 1 \leq x < 2(n + 1),
\end{cases}
\]

where \( n \) is an integer. We call the second example the Tent system as \( f \) has the tent map like shape. Numerical integration of Eq. (1) was performed by the Runge-Kutta scheme with the step size \( \Delta t = 0.001 \). Figures 1 (a) and (b) show diagrams of the LCLE \( \lambda \) in the \((D,\alpha)\) plane for the time-delay systems (25) and (26), respectively. The curves of \( \lambda = 0 \) are shown by solid line and the regions of positive and negative LCLE are also indicated. The parameter \( \beta \) are fixed as \( \beta = 5 \) in both of the examples.

In both of the figures, the LCLE is negative for any \( D \) in small \( \alpha \) regions. This observation is in agreement with the analytical criterion given by Eq. (6). The boundary lines associated with Eq. (6) are shown by dashed line. The regions under the dashed lines meet the inequality (6). It can be confirmed that these regions are located inside the negative LCLE regions obtained numerically.

Figures 1 (a) and (b) show that the LCLE is positive for \( \alpha \) larger than some critical values when \( D = 0 \). This means that the noise-free systems are chaotic for those values of \( \alpha \). The critical values in \( \alpha \) are found about 2.6 and 1.4 in Figs. 1 (a) and (b), respectively. The numerical results show that for any \( \alpha \) the LCLE is negative and thus the IC occurs when the noise intensity \( D \) is large. Large noise intensity gives rise to the IC even if the noise-free system is chaotic. This numerical observation coincides with the theoretical result.

5. Conclusions

In conclusion, we studied the IC in a class of one-dimensional time-delay systems. It was analytically shown that large noise intensity \( D \) causes the IC for any given \( \alpha \) in the time-delay systems satisfying the assumptions (i)-(iii), provided that \( e^{-D} \ll 1 \). Moreover, it was also analytically shown that the IC occurs for any \( D \) when \( \alpha \) is small. These analytical results were confirmed by numerical experiments.

References

Consistency and ICA of Solvable Chaotic Maps with Non-Gaussian Stable Distributions

Ken Umeno†, Atsushi Uchida‡, and Minghui Kao*

† National Institute of Information and Communications Technology
4-2-1 Koganei, Tokyo 184-8795, Japan
‡Department of Electronics and Computer Systems, Takushoku University
815-1, Tatamachim Hachioji-shi, Tokyo 193-0985, Japan
*GITS, Waseda University
Shinjuku, Tokyo, Japan
Email: umeno@nict.go.jp, a-uchida@es.takushoku-u.ac.jp,kao@akane.waseda.jp

Abstract– A rigorously solvable model for consistency of a discrete-time nonlinear system’s response to a repeated chaotic drive signals is proposed with exactly solvable chaotic mapping with non-Gaussian stable ergodic invariant measures. We show that an exact condition for the consistency is obtained by stable property of ergodic invariant density. Furthermore, we show that this rigorous result is in good agreement with numerical simulations.

1. Introduction

Many nonlinear system have consistent outputs when driven by a repeated external signal. Thus, the concept of consistency defined as the reproducibility of response waveforms in a nonlinear dynamical system driven by a signal, starting from different initial conditions of the system could be widely applicable to test complex nonlinear phenomena[1]. Here we propose a discrete time nonlinear system driven by a repeated external chaotic signal given by a chaotic mapping with non-Gaussian stable density.

2. Solvable Chaos and Levy’s Chaos

2.1. Solvable Chaos

Solvable chaos is defined as a chaotic dynamical system with exact ergodic density function such that exact probabilistic characteristic is exactly obtained [2]. Such solvable chaos model can be given by an addition theorem of elliptic function and sine function.

2.2. Levy’s Chaos

Levy’s chaos is defined as a chaotic dynamical system with non-Gaussian density function such that whose superposition would converge to Levy’s stable law by the generalized central limit theorem. In 1998, a general model for generating Levy’s chaos was obtained by one of the present authors [3].

2.3. Cauchy Chaos

Cauchy Chaos is defined by a class of dynamical systems whose ergodic invariant density functions are given by the Cauchy distribution over the real. Since, the Cauchy distribution is a special class of Levy’s stable distribution, Cauchy Chaos can be regarded as a special class of Levy’s Chaos. Such Cauchy chaos is illustrated by the following example [3]:

\[ x[n+1]=1/2(x[n]-1/x[n])=f[x[n]]. \] (1)

This dynamical system has an exact density function given by

\[ \rho (x)=1/\pi (1+x^2). \] (2)

3. Solvable Model of Consistency

We consider the consistency by the Cauchy Chaos in the previous section. Our model is defined as follows:

\[ x[n+1]=f[x[n]]+\epsilon D[n] \] (3)

\[ y[n+1]=f[y[n]]+\epsilon D[n] \] (4)

\[ D[n+1]=f[D[n]] \] (5),

where the function \( f[\cdot] \) is defined by (1), \( D[j] \) is a drive signal and \( \epsilon \) measures the relative strength of drive signal. The initial conditions \( x[0] \) and \( y[0] \) are different each other.

By the stable property of characteristic function of stable distributions with superposed variable such as Eq. (3) and Eq. (4), the superposed distribution function of Eq. (3) and Eq. (4) is given by

\[ \rho (x)=a/\pi (a^2+x^2), \] (6)

where

\[ a=1+\epsilon. \] (7)
Thus, the conditional Lyapunov exponent $\lambda_c$ of (3) and (4) is given by

$$\lambda_c = \int \{\log(1/2) + \log(1+1/x^2)\} \rho_s(x) \, dx \quad (8)$$

where the integral is integrated over the infinite interval from $-\infty$ to $\infty$ and $\rho_s(x)$ is the superposed density function given by Eq. (6).

Remarkably, $\lambda_c$ can be exactly computed as

$$\lambda_c = 1/a \{ \tanh^{-1}(1/a) + \log[1 - 1/a^2] \} - \log 2 \quad (9)$$

where $a = 1 + \varepsilon$ is defined by a drive signal amplitude $\varepsilon$ as given by Eq. (7).

If $\lambda_c$ is less than 0, the consistency is predicted to occur. On the other hand, if $\lambda_c$ is greater than 0, the consistency is predicted to not occur. Thus, the sign of $\lambda_c$ is important.

Figure 1 shows the relation between $\varepsilon = 1/a \{ \tanh^{-1}(1/a) + \log[1 - 1/a^2] \} - \pi \log 2$ and $a$.

Figure 2 shows that a properly normalized cross correlation is converged to unity (consistent) after $\varepsilon \geq 1.4$ in our numerical simulations [4].

In Fig. 1, it is shown that $a = 2.4$ is a critical value for the consistency. Thus $\varepsilon = 1.4$ can be predicted to be a critical drive signal strength parameter for the consistency in this model.

By solving the equation

$$\lambda_c = 1/a \{ \tanh^{-1}(1/a) + \log[1 - 1/a^2] \} - \log 2 = 0 \quad (10)$$

we can obtain the exact critical drive signal strength parameter $\varepsilon_c$.

This value was greatly consistent with our numerical simulations initially reported in Ref. [4] where the cross correlation of the two response signals as a function of the amplitude of the drive signal is converged to unity at $\varepsilon = 1.4$. 

Fig. 1. Rescaled Conditional Lyapunov exponent $\pi \lambda_c$.

Fig. 3. Non-Gaussian Chaotic Time Series at $\varepsilon = 0.5$ and $\varepsilon = 1$. 

Fig. 2. Cross-correlation versus Drive Signal Amplitude.
In Fig. 3, time series of \( x(n) \) and \( y(n) \) is plotted for \( \lambda = 0.5 \) and \( \lambda = 1 \). They are chaotic as predicted by the positive conditional Lyapunov exponent given in Eq. (9).

4. Consistency and ICA

Consistency is a strongly dependent phenomenon in nonlinear systems characterized negative conditional Lyapunov exponent with driven signals. Conversely, independent features of nonlinear dynamical variable can also characterize this consistency. In this respect, ICA (independent component analysis) for this nonlinear phenomena can measure the degree of consistency. ICA is a strong signal separation method for mixed chaotic signals with future mobile communications [5]. Thus, it is of interest to see more general concept of consistency can be defined by the generalized linear response of ICA framework. Let consider a more generalized model of consistency as follows:

\[
\begin{align*}
x[n+1] &= f(x[n]) + e_1D_1[n] + e_2D_2[n] + \ldots + e_mD_m[n] \\
y[n+1] &= f(y[n]) + h_1D_1[n] + h_2D_2[n] + \ldots + h_mD_m[n] \\
D_1[n+1] &= f(D_1[n]) \\
&\vdots \\
D_m[n+1] &= f(D_m[n]).
\end{align*}
\]

If the initial conditions of Eq. (11)-Eq. (14) are different each other, a dynamical drive signals \( D_1[n], D_2[n], \ldots, D_m[n] \) imputed for \( x(n) \) and \( y(n) \) could have a mixing property [3]. In other words, those dynamical variable can be considered as independent. Thus, the characteristic functions of the distribution of Eq. (11) and Eq. (12) can be written as a product of the characteristic functions of the distributions of \( x[n] \) (or \( y[n] \)), \( D_1[n], D_2[n], \ldots, D_m[n] \). Thus, the superimposed distribution functions of Eq. (11) and Eq. (12) can be written by the following formula:

\[
\begin{align*}
\rho(x) &= a/(\sqrt{\Pi(\pi a^2+x^2)}), \\
\rho(y) &= b/(\sqrt{\Pi(\pi b^2+y^2)}),
\end{align*}
\]

where

\[
\begin{align*}
a &= 1 + |e_1| + |e_2| + \ldots + |e_m|, \\
b &= 1 + |h_1| + |h_2| + \ldots + |h_m|.
\end{align*}
\]

The consistency can occur for this model if the following condition is satisfied:

\[
a = 1 + |e_1| + |e_2| + \ldots + |e_m| = b = 1 + |h_1| + |h_2| + \ldots + |h_m|.
\]

In this case, the local Lyapunov exponent \( \lambda c \) is given by Eq. (9). In Eq. (9), the local Lyapunov exponent \( \lambda c \) is a decreasing function with the drive signal amplitude \( a \) is getting larger.
Thus, we can see that adding independent drive signals can also enhance consistency. It is a well-known fact that ICA of many independent variables is difficult and cannot separate signals due to the central limit theorem assuring the universal convergence to the Gaussian distribution of the sums of independent variables. This fact is consistent with our present observation as above saying that consistency overwhelms independent behavior when adding many independent drive signals. Thus, we predict that consistency can be understood and characterized to some extent in the framework of generalized central limit theorem for nonlinear systems with many variables. In this sense, the role of stable distribution such as the Cauchy distribution is very important for further understanding consistency because a superposition process of the variables does not lose the stable characteristic of the distribution functions. The effect of our generalization to many drive signals in our present generalized framework in Eq. (11) and Eq. (12) can be seen in a system whose density function is not a stable law where a superposition effect of many small independent drive signals enhances the convergence to stable law. Such a system is illustrated by the following chaotic map [3]:

\[ f(x_i) = \frac{1}{2} \left( x_i \right) - \frac{1}{|x_i|} \left| x_i \right|^{\alpha} \text{sgn}(x_i - \frac{1}{x_i}) \]

where the ergodic invariant density

\[ \rho_{eq}(x) = (\alpha/\pi) \left| x \right|^{-\alpha-1} \left( 1 + \left| x \right|^{2\alpha} \right) \]

and \( \alpha \) is the Levy’s index of stable law. It is noted that if \( \alpha = 1 \), it is exactly the Cauchy distribution. Thus, the universal feature of the generalized central limit theorem for such generalized chaotic maps is expected to reveal the universal feature of consistency by considering this kind of generalization of model of consistency.

5. Conclusion

We obtain the exact solvable model for consistency with exactly solvable models of chaos with non-Gaussian density functions. By solvable, it means that consistency–non-consistency behavior is exactly characterized by the parameter of the amplitude of drive signals as in Eq. (10) in the same spirit of [6]. Furthermore, we discuss the relation between consistency and ICA by considering a generalized model of consistency with many independent drive signals. It turns out that the universal characteristic of consistency and ICA can be characterized by the universal feature of generalized central limit theorem.

References


Acknowledgements

One of the authors (K.U.) would like to thank Mr. Masahiko Kuraya for discussion and his assistance with numerical simulations. This work is financially supported by NEDO Grant for Industrial Technology Research (FY2005-2).
Reproducibility of limit-cycle oscillators induced by random impulses

Kensuke Arai and Hiroya Nakao
Department of Physics, Kyoto University, Kyoto 606-8502, Japan
Email: arai@ton.scphys.kyoto-u.ac.jp, nakao@ton.scphys.kyoto-u.ac.jp

Abstract
We argue that the reliability of a limit-cycle oscillator generally improves when it is driven by a random impulsive input. By reducing the dynamics of the oscillator to a stochastic phase equation, we argue generally that the orbit of a limit-cycle oscillator is statistically stabilized against phase disturbances when it is driven by a weak impulsive input regardless of its details, leading to an improvement in its reproducibility. We demonstrate our theoretical results by numerical simulations and experiments with a simple electrical oscillator.

1. Introduction
It is now well established that mutually interacting limit-cycle oscillators synchronize with each other [1]. Recently, it has been demonstrated that nonlinear oscillators can synchronize with each other through common fluctuating inputs, even in the absence of direct mutual interactions. Such noise-induced synchronization of an ensemble of nonlinear oscillators can also be interpreted as noise-induced improvement in the reproducibility of the orbit of a single oscillator, because repeated experiments with a single oscillator using the same input signal is equivalent to a single experiment with multiple oscillators using a common input signal. Noise-induced synchronization, reproducibility, or “consistency” as termed by some authors, is widely observed in experimental systems ranging from lasers to neurons [2, 3], where the dynamics of each element can also be chaotic or stochastic.

For example, in electrophysiological experiments using a single neuron from slice preparations of rat neocortex or olfactory bulb [3], it is observed that the neuron evokes different spike sequences from trial to trial when it is driven by a constant input current, whereas it evokes mostly the same spike sequences over the trials when it is driven by a fluctuating input current. This phenomenon can be interpreted as a noise-induced improvement in the reproducibility of the orbit of a limit-cycle oscillator induced by a fluctuating input signal.

On the theoretical aspect, after several pioneering works [4], Teramae & Tanaka and Goldobin & Pikovsky [5] adopted the phase reduction method [1] and proved generally that uncoupled limit-cycle oscillators always synchronize with each other when they are driven by a common weak Gaussian white noise. Using a similar idea, we also argued generally that uncoupled limit-cycle oscillators can synchronize with each other when they are driven by random impulsive, telegraphic, or piecewise-constant input signals [6].

In this paper, we adopt the theory of Poisson driven Markov process [7] to analyze the case of impulsive input signals. We generally prove that the reproducibility of an oscillator orbit always improves for sufficiently weak Poisson impulses. We also argue that when the impulses are not weak, the oscillators can also undergo desynchronization rather than synchronization. We demonstrate our theoretical results experimentally by observing the effects of Poisson impulses applied to a simple electrical circuit exhibiting limit-cycle oscillations.

2. Numerical Example
As an example, we first present the result of direct numerical simulations using the FitzHugh-Nagumo (FN) neural oscillator. The FN oscillator is described by

\begin{align}
\dot{u}(t) &= \epsilon(v + c - du), \\
\dot{v}(t) &= v - v^3/3 - u + I_0 + I(t) + \xi(t),
\end{align}

where \(v(t)\) represents the membrane potential of the neuron at time \(t\), \(u(t)\) the state of the ion channels in a reduced form, \(\epsilon, c, a, d\) are parameters, \(I_0\) and \(I(t)\) represent the constant component and the fluctuating component of the input current. \(\xi(t)\) represents a weak Gaussian white noise specified by \(\langle \xi(t) \rangle = 0\) and \(\langle \xi(t)\xi(s) \rangle = D\delta(t - s)\), which incorporates the effect of various fluctuations. When \(\xi(t)\) and \(I(t)\) are absent and \(I_0\) is fixed at a constant value in some appropriate range, this model exhibits a typical limit-cycle oscillation, which corresponds to the periodic spiking of the neuron. We fix the parameters at \(\epsilon = 0.08, c = 0.7, d = 0.8, I_0 = 0.8,\) and \(D = 0.0001\) in the following.

Figure 1 displays the result of 50 repeated numerical simulations of Eq. (1), where the spiking times of the oscillator (the point at which \(v\) changes its sign from negative to positive) are plotted by dots. In Fig. 1(a), the oscillator is driven only by the constant input current \(I_0\). Due to the weak noise \(\xi(t)\) applied independently at every trial, the spiking times of the oscillators are considerably scattered. This is due to the neutral stability of the limit cycle in the phase direction. The orbital component of the perturbations do not decay and gradually accumulate, which results
in phase diffusion. Figure 1(b) displays the result of 50 repeated simulations in which the oscillators are driven not only by the constant current $I_0$ but also by random Poisson impulses $I(t)$ given by
\[ I(t) = \sum_{n=1}^{N(t)} I_0 \delta(t - t_n), \] (2)
where $\{t_1, t_2, \cdots\}$ represent the generation times of the impulses, and $\{I_1, I_2, \cdots\}$ the intensity of the impulses. We set the mean interval between the impulses at $\tau = 10$ and assume that the intensity $I_0$ of the impulse takes $\pm 0.5$ with equal probability. In this case, the spiking times of the oscillator are reliably reproduced over the trials after an initial transient, even under the effect of the weak independent noises. We analyze the mechanism leading to such behavior in the next section.

3. Phase Reduction Analysis

The improvement in reproducibility is due to the statistical stabilization of perturbations in the phase direction induced by random impulses. We formulate this fact using stochastic differential equation in this section. Our model is generally described by the following random dynamical system:
\[ \dot{X}(t) = F(X; I_0) + I(t), \quad I(t) = \sum_{n=1}^{N(t)} I_0 \delta(t - t_n), \] (3)
where $X$ represents the dynamical variable of the oscillator, $F$ its dynamics, $I_0$ the constant component of the external input, and $I(t)$ the random impulsive input. We assume that the impulses are generated by a Poisson process of rate $\lambda$ (so that the mean interval is $\tau = \lambda^{-1}$). We denote the total number of impulses generated up to time $t$ by $N(t)$, and the generation times and intensities of the $n$-th impulse by $\{t_n, I_n\}$. Each impulse intensity $I_n$ is chosen randomly from a probability density function $Q(I)$. We omit the weak independent noise in the following analysis, which is not important in the linear stability analysis. We assume that the system has a single limit cycle $X_0(t)$ when only the constant input $I_0$ is given, and also that most initial conditions in the phase space are eventually attracted to this limit cycle.

When $\lambda$ is small, the mean interval between the impulses becomes long, so that the oscillator almost always receives only the constant input $I_0$. The orbit is kicked off from the limit cycle occasionally by an impulse, but it returns to the limit cycle sufficiently quickly before the arrival of the next impulse. In such a situation, we can reduce the evolution equation (3), which generally contains multiple variables, to a simplified equation of the scalar phase variable only. In our previous paper [6], we formulated such phase reduction using random phase maps. Here, we adopt the methods of stochastic differential equations [7] for this purpose.

We first rewrite Eq. (3) as a stochastic differential equation driven by the Poisson impulses as [7]
\[ dX(t) = F(X; I_0)dt + \int I M(dt, dI). \] (4)
Here, $M(dt, dI)$ denotes the Poisson random measure [7], which represents the number of impulses generated in the time interval $[t, t + dt]$ whose intensity is in the range $[I, I + dI]$. Its expectation value is given by $E[M(dt, dI)] = \lambda Q(I)dt dI$. We then introduce a phase variable $\theta(X) \in [0, 1]$ along the limit cycle $X_0(t)$ corresponding to the constant input $I_0$, which increases with constant angular velocity $\omega$. We can extend this definition of the phase variable to the whole of phase-space (except phase singular points) by assigning the same phase value to the set of points that eventually converge to the same point on the limit cycle [1].

By applying the (generalized) Ito formula for the Poisson driven Markov process [7] to Eq. (4), we obtain
\[ d\theta(t) = \omega dt + \int \left[ \theta(X(t) + I) - \theta(X(t)) \right] M(dt, dI). \] (5)
Furthermore, by assuming that the orbit is (almost) always on the limit cycle when it receives an impulse, we approximate the $X(t)$ in the above equations by $X_0(\theta(t))$. We then obtain a closed phase equation for $\theta(t)$,
\[ d\theta(t) = \omega dt + \int G(\theta(t), I) M(dt, dI), \] (6)
where the function $G(\theta, I)$ is defined as
\[ G(\theta, I) = \theta(X_0(\theta) + I) - \theta. \] (7)
$G(\theta, I)$ is a periodic function in $\theta$ representing the change in the phase of the orbit after receiving an impulse of intensity $I$ at the point $X_0(\theta)$ on the limit cycle. We refer to this function as “phase map” hereafter.
Now we consider the statistical stability of the phase against small perturbations. A linearized equation for the evolution of a small perturbation $\psi(t)$ to the original phase $\theta(t)$ is obtained from Eq. (6) as

$$d\psi(t) = \int G'(\theta(t), I)\psi(t)M(dt, dI),$$

where $'$ denotes differential by $\theta$. By applying the Ito formula to this equation, we obtain the following equation for the logarithm of the absolute phase perturbation $\ln |\psi(t)|$:

$$d\ln |\psi(t)| = \int \ln |1 + G'(\theta(t), I)| M(dt, dI).$$

By taking the expectation, we obtain

$$E[d\ln |\psi(t)|] = \Lambda dt,$$

where we defined the Lyapunov exponent $\Lambda$ that quantifies the mean growth rate of the small perturbation as

$$\Lambda = E[\ln |1 + G'(\theta(t), I)|].$$

Here, $P(\theta)$ is the stationary probability density of the phase $\theta$. When this $\Lambda$ is negative, the phase perturbation decays exponentially, resulting in the improvement of reproducibility of the limit-cycle oscillator.

Particularly, when the intensities of the external impulses are sufficiently small, we can generally argue for the negativity of $\Lambda$ as follows. In this case, $G(\theta, I)$ can be approximated as

$$G(\theta, I) \approx Z(\theta) \cdot I,$$

where $Z(\theta) := \nabla_\theta \psi(X)|_{X = X_\theta}$ is the conventional phase sensitivity function [1]. Since $Z(\theta)$ is a smooth periodic function in $\theta$, $G(\theta, I)$ is also smooth and does not fluctuate largely when $|I|$ is small. In such a situation, $1 + G'(\theta, I)$ is always positive. Also, it is physically apparent that the probability density function of the phase $\theta(t)$ becomes almost uniform,

$$P(\theta) \approx 1,$$

when the impulses are weak (this can also be shown analytically using the Master or the Frobenius-Perron equations).

Now, by using the periodicity of $G(\theta, I)$ in $\theta$, we can generally prove the negativity of $\Lambda$ as follows:

$$\Lambda = \lambda \int dQ(I) \int d\theta \ln |1 + G'(\theta, I)|$$

$$\leq \lambda \int dQ(I) \int d\theta G'(\theta, I)$$

$$= \lambda \int dQ(I)[G(1, I) - G(0, I)]$$

$$= 0.$$  

The equality holds only in the trivial case of the constant $G(\theta, I)$. Thus, we can prove that $\Lambda$ is always less than 0 for a general class of limit-cycle oscillators regardless of their details when the random impulses are sufficiently weak. Therefore, it is expected that the reproducibility of a limit-cycle oscillator always improves when it is driven by sufficiently weak random impulses. In the next section, we will demonstrate this fact using an electric circuit receiving random impulses.

Figure 2: Phase maps of the FN model for several values of the impulse intensity $I$.

Let us examine the phase maps of the FN oscillator now. Figure 2 displays the phase maps $G(\theta, I)$ obtained for several values of the impulse intensity $I$. When $|I|$ is small, the amplitude of the corresponding phase maps are also small, so that $\Lambda$ is negative from the above argument. Thus, the reproducibility improves by the external impulses as demonstrated in Fig. 1(b). In contrast, when $|I|$ is large, the amplitude of the corresponding phase maps can also become large and fluctuates strongly (Eq.(12) does not hold for large impulses). $\Lambda$ can be positive in such cases, leading to impulse-induced desynchronization.

4. Electric Circuit Experiments

In this section, we demonstrate the phenomenon of impulse-induced reproducibility in a real experimental system using a very simple electric circuit.

Our electric circuit was originally designed for flashing an LED periodically (Fig. 3(a)). We use a computer equipped with an AD/DA-converter board (CONTEC ADA16-8/2) to generate an output signal (the impulses), and to measure the voltages $V_1$ and $V_2$ at 2 separate locations in the circuit. Using a MOSFET, voltage impulses from the AD/DA board was used to briefly short circuit a given section of the oscillator. The strength of the impulse delivered was varied by adding a variable resistor in series with the MOSFET. Changing the location in the circuit in which to deliver the impulse naturally changes the response of the circuit to a given impulse.
Figure 3(b) displays typical limit cycle of the electric circuit, which is qualitatively similar to the FN oscillator we treated above. Figure 3(c) shows experimentally observed phase maps obtained for two different intensities of the impulses (i.e. the magnitude of the series resistor). With these intensities of the impulses, the function $G(\theta, I)$ is sufficiently smooth, and the absolute value of $G'(\theta, I)$ is always smaller than 1. Therefore, we anticipate impulse-induced reproducibility in this case. The results are shown in Figs. 3(d) and (e), where two time sequences of $V_1$ measured in two different experimental trials, and and their spiking times are plotted. Under the effect of random impulses, the two temporal sequences behave quite similarly. In contrast, without impulses, the two temporal sequences behave quite differently due to the internal noises of the electric circuit. Thus, we see that noise-induced synchronization can easily be observed in our simple experiment.

Figure 3: Results of the circuit experiment. (a) circuit diagram, (b) limit cycle, (c) phase maps $G(\theta, I)$ obtained using two intensities of the impulses, (d) superposition of two measured temporal sequences of $V_1$, and (e) spiking times.

Though noise-induced synchronization has already been observed in a wide class of systems, we believe that carefully controlled experiments using simple experimental systems can also be useful in deepening our insight in this phenomenon. Specifically, the experimental measurement of the phase map, which is rather difficult in more complicated and widely fluctuating systems such as neurons, can be accomplished simply and unambiguously in a simple system such as this. Our results reported here are only preliminarily. More detailed quantitative analysis will be reported in the future.

5. Summary

We proved theoretically that the reproducibility of noisy limit-cycle oscillators generally improves when driven by random impulses, and experimentally confirmed this theoretical prediction using an electric circuit. Our theory holds for any limit-cycle oscillators that satisfy our assumptions irrespective of their detailed structures. We thus expect that synchronization or reproducibility of randomly driven nonlinear oscillators caused by this mechanism can be observed in various natural phenomena.

Acknowledgments

The authors thank Y. Kuramoto, J. Teramae, Y. Tsubo, K. Nagai, and A. Uchida for useful comments.

References


Multiple basins of consistency in a driven Mackey-Glass electronic circuit

Hoipang Yip, Satoshi Sano, Atsushi Uchida, and Shigeru Yoshimori

†Department of Electronics and Computer Systems, Takushoku University
815-1 Tatemachi, Hachioji, Tokyo 193-0985, Japan
Email: k_sinbad@hotmail.com

Abstract—We experimentally observed consistency in a Mackey-Glass electronic circuit (MGC) driven by a MGC-produced chaos signal and a colored-noise signal. We discovered that two types of consistent response signals are observed in the case of colored-noise drive, which is called “multiple basins of consistency”.

1. Introduction

As for many nonlinear systems, consistent behavior is shown to a repeated input signal. “Consistency” indicates “reproducibility of the output of a nonlinear system repeatedly driven by a certain signal. Consistency has been considered as a common concept observed in various interdisciplinary fields.

The experimental observation of consistency has been reported in a microchip solid-state laser [1]. Some numerical simulations of consistency have been investigated [2, 3]. However, no experimental observation has been reported in other nonlinear systems, such as electronic circuits.

In this letter, we experimentally observe consistency in a Mackey-Glass electronic circuit driven by a chaos signal and a colored noise signal. We quantitatively investigate the degree of consistency when the amplitude of driving signal is changed.

2. Mackey-Glass electronic circuit

Leukocyte productivity dynamics of human beings is as the origin of Mackey-Glass model [4]. The model equation is as follows:

\[
\frac{dx(t)}{dt} = \frac{ax(t - \tau)}{1 + cx^n(t - \tau)} - bx(t) \quad \cdots \cdots (1)
\]

This equation contains a time delay \(\tau\) and a nonlinear function, and generates a very complex chaotic waveform with high dimensionality.

This model is implemented in an electronic circuit. The circuit diagram is shown in Fig. 1. The circuit consists of two field effect transistors (2SJ103 and 2SK30A) for the nonlinear part, three amplifiers, a RC filter, and a delay line. Fifty pairs of capacitors (10 nF) and inductors (4.7 mH) are combined to generate time delay. The total delay time is 318 µs which corresponds to a half period of chaotic waveforms.

3. Experiment

3.1. Experiment of consistency with chaos drive signal

The experimental setup for consistency is shown in Fig. 2. The parameters of the Mackey-Glass electronic circuit are set to generate chaos and the chaos signal is detected by an oscilloscope. This chaos signal is recorded by a computer, and used as a signal for the drive. The chaos signal recorded in the computer is generated from a function generator and added with the feedback signal from the MGC, then re-injected to the MGC to observe consistency. The recorded chaos drive signal is sent to the MGC twice and two response waveforms are detected.
Figure 3 shows the temporal waveforms of chaos drive signal and two response waveforms and the correlation plots between the two response waveforms. The two response waveforms obtained from the same drive signal are almost the same, i.e., the consistency exists. Linear cross correlation is obtained between the two response waveforms as shown in Fig. 3(b).

The ratio of the drive signal and the feedback signal is changed, and the presence of consistency was evaluated by using the cross correlation function of two response signals. The result is shown in Fig. 4. Cross correlation is around 1 as the drive signal ratio is increased. We found that consistency is obtained by driving MGC with a large drive signal.

### 3.2. Experiment of consistency with colored-noise drive signal

Next, a colored noise signal is generated in a computer as a drive signal. This drive signal is sent four times to the Mackey-Glass electronic circuit. Figure 5 shows the temporal waveforms of the colored-noise drive signal and the four response waveforms. Consistency is observed between Response 1 and 3, and between Response 2 and 4.

However, Response 1 and 2 (also Response 3 and 4) are not consistent waveforms. Figure 6 shows the correlation plots between two of four response waveforms. It is found that Response 1 and 3 are consistent, and Response 2 and 4 are consistent. The other combinations are not consistent. i.e., there are two types of consistent response waveforms as shown in Figs. 5 and 6.
We calculate the cross correlation between two of the four response waveforms and the average value is plotted as a function of the drive signal ratio as shown in Fig. 7. The error bars shows the maximum and minimum values of the cross correlation. We found that the average cross correlation function is not kept constant to 1 at large \( \kappa \) (see error bars in Fig.7). At the region shown in a large error bar, two pairs of consistent waveforms are obtained as shown in Fig. 6. This result shows that there are two basins of consistency depending on the initial conditions.
This is a concept similar to the multi-stability where two or more attractors exist. The concept of “multiple basins of consistency” may be a general concept that can be observed in many nonlinear dynamical systems.

4. Conclusion

In this study, we have experimentally observed consistency in the Mackey-Glass electronic circuit driven by chaos and colored-noise. Consistency is obtained in the case of both chaos and colored-noise drive signals with large amplitude. We have observed two types of response waveforms when using a colored-noise drive signal. This observation leads to a new universal concept of “multiple basins of consistency” in non-autonomous nonlinear dynamical systems.

References

Consistency Caused by Various Types of Signals

GOTO Shin-itiro and YOSHIMURA Kazuyuki

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
Email: {sgoto,kazuyuki}@cslab.kecl.ntt.co.jp

Abstract—We study the synchronization of dynamical systems induced by a common additional external signal. In particular, we consider the special case that the external input signal is generated by a linear second-order differential equation forced by Gaussian white noise. For the case that noise-free systems are chaotic, we show that forced systems can be in resonance with the input signals. In addition, the resonance frequency is the same as the frequency of the noise-free system under study. On the other hand, for the case that noise-free systems are non-chaotic, we show that forced systems cannot be in resonance with the input signals.

1. Introduction

Many dissipative dynamical systems have an ability to generate a consistent output when driven by an external signal. Consistency [1] is defined as the reproductibility of a response waveform in a dynamical system driven by a signal, starting from different initial conditions of the system.

In experiments, consistency has been studied in dynamical systems, such as lasers[1], neural network [2] and ecological systems[3]. On the other hand, it has theoretically been shown that consistency occurs in a wide class of limit cycle oscillators[4]. In addition, consistency has been demonstrated in chaotic systems numerically[5].

In the studies of consistency, the case of Gaussian white noise input has been intensively investigated. However, there are a few studies on consistency in systems forced by colored noises. In the case when a colored noise is adopted, the Ornstein-Uhlenbeck noise, which is generated by the first order differential equation forced by Gaussian white noise, has only been considered[6, 7].

In the real world, there are various types of waveforms of the input signal to a system. We are then interested in whether or not the input to a system affects the nature of consistency. In particular, we would like to know what kinds of inputs induce or enhance consistency in a system, beyond the previous studies.

In this paper, we study consistency for systems in the special case that input signals are generated by a linear second-order differential equation forced by Gaussian white noise. According to Ref.[6], we classify systems into two categories: chaotic and non-chaotic in the noise-free dynamics. After giving the definition of consistency and a method to measure the consistency in a system, we numerically show that forced systems exhibiting consistency, which are chaotic in the noise-free cases, can be in resonance with the input signals. The resonance frequency is given by the typical frequency of the noise-free system under study.

2. Input Signal and Measurement of Consistency

To discuss consistency, we consider the following system:

\[ \dot{x} = F(x) + C \varepsilon(t), \]  

where \( x(t) \) is real, \( \dot{x} \) denotes the differentiation of \( x \) with respect to time, \( F \) is a real function, \( C \) is a real constant vector, and \( \varepsilon(t) \in \mathbb{R} \) is a common input explained in the next subsection. We study cases in which noise-free systems, \( \dot{x} = F(x) \), are chaotic and non-chaotic.

2.1. Input Signal

We prepare the input signal \( \varepsilon(t) \) using a second-order linear differential equation forced by Gaussian white noise. That is, the input signal \( \varepsilon(t) \) which we use in this paper is obtained through

\[ \dot{\varepsilon} + \gamma \varepsilon + \gamma^2 \omega_0^2 \varepsilon = \gamma^2 \eta(t), \]

\[ \langle \eta(t) \rangle = 0, \quad \langle \eta(t)\eta(t') \rangle = 2D \delta(t - t'). \]

Here \( \gamma, \omega_0 \) are positive constants, \( \eta(t) \) is Gaussian white noise, \( D \) is a positive constant, \( \delta(t) \) is Dirac’s delta function, and \( \langle \cdots \rangle \) denotes averaging over the realization of \( \eta \). The autocorrelation function of the signal \( \varepsilon(t) \) for \( \omega_0 > 1/2 \) and \( t > 0 \), is obtained as[8]

\[ \phi_\varepsilon(t) := \langle \varepsilon(t)\varepsilon(0) \rangle = \frac{\gamma D}{\omega_0^2} \left\{ \cos(\omega_1 t) + \frac{\sin(\gamma \omega_1 t)}{2\omega_1} \right\} e^{-\gamma t/2}, \]

where \( \omega_1 := \sqrt{\omega_0^2 - 1/4} \). In the other parameter regime, \( 0 \leq \omega_0 \leq 1/2 \), the autocorrelation function of \( \varepsilon(t) \) is over-damped, and in consequence, the autocorrelation of \( \varepsilon(t) \) could be expected to be similar...
to that of the Ornstein-Uhlenbeck noise. Then we only concentrate on the regime \( \omega_0 > 1/2 \), in this paper. We define the strength \( \sigma \) of the input signal by its standard deviation. From Eq. (4), \( \sigma \) is given by \( \sigma := \sqrt{\phi_\Sigma(0)} = \sqrt{\gamma D/\omega_0} \). The power spectrum of \( \epsilon(t) \) is calculated through Eq. (4) [8],

\[
I_\epsilon(\omega) = \frac{D^2}{\pi} \frac{1}{(\gamma^2 \omega_0^2 - \omega^2)^2 + \gamma^2 \omega^2}. \tag{5}
\]

It is noted here that the maximum of \( I_\epsilon(\omega) \) is obtained at \( \omega = \gamma \omega_0 \), and that \( \epsilon(t) \) become Gaussian white noise with an appropriate factor as the limit \( \gamma \to \infty \).

To compare the previous studies [6, 7] with ours, we briefly give the nature of the Ornstein-Uhlenbeck noise. The Ornstein-Uhlenbeck noise \( \epsilon(t) \) is generated by

\[
\dot{\epsilon} = -\gamma \epsilon + \eta(t),
\]

where \( \gamma \) is a real parameter and \( \eta(t) \) is Gaussian white noise given by Eq. (3). The autocorrelation function \( \phi_\epsilon(t) := \langle \epsilon(t) \epsilon(0) \rangle \) and the power spectrum \( I_\epsilon(\omega) \) are calculated as [8]

\[
\phi_\epsilon(t) = D \gamma \exp(-\gamma |t|), \quad I_\epsilon(\omega) = \frac{D^2 \gamma^2}{\omega^2 + \gamma^2}. \tag{6}
\]

From the expression of \( I_\epsilon(\omega) \), the maximum of \( I_\epsilon(\omega) \) is obtained at \( \omega = 0 \).

### 2.2. Measurement of Consistency

In this subsection, we first quantitatively define the consistency for the system (1). Second, we give a method to detect the consistency and to measure it.

Suppose two trajectories \( x_1(t) \) and \( x_2(t) \) in Eq. (1) with different initial conditions for the same realization of \( \epsilon(t) \). The consistency is said to occur when

\[
\lim_{t \to \infty} \| x_2(t) - x_1(t) \| = 0,
\]

for any initial conditions. Here \( \| \cdots \| \) expresses Euclidean norm of the argument.

To measure consistency, we introduce the largest conditional Lyapunov exponent. A linear stability analysis for the system (1) is performed with a small derivation \( \delta x(t) := x_2(t) - x_1(t) \) from the state \( x(t) = x_2(t) = x_1(t) \), whose stability is governed by

\[
\dot{\delta x} = (DF)(x) \delta x, \tag{7}
\]

where \( (DF)(x) \) is the Jacobian matrix of \( F \) evaluated at \( x \). It is noted here that \( \eta(t) \) affects not only the time-evolution of \( x(t) \) but also the stability (7). The largest conditional Lyapunov exponent is defined by

\[
\lambda := \lim_{t \to \infty} \frac{1}{t} \ln \frac{\| \delta x(t) \|}{\| \delta x(0) \|},
\]

Figure 1: For the noisy Lorenz system with \( a = 10, b = 28 \) and \( c = 8/3 \). (a): The threshold noise amplitude \( \sigma_c \) vs. \( \gamma \), \( (\omega_0 = 1) \). (b): The contour plot of the largest conditional Lyapunov exponent in the \( \gamma \omega_0 - \gamma \) plane, \( (\sigma = 70) \).

Clearly \( \lambda \) expresses the stability of the state \( x_1(t) = x_2(t) \). In this paper, the consistency is said to occur when \( \lambda < 0 \). Furthermore, when \( \lambda < 0 \), the time to reach the state \( \| x_2(t) - x_1(t) \| = 0 \) becomes shorter as \(|\lambda|\) increases, due to the relation \( \delta x(t) \sim \delta x(0) \exp(\lambda t), (t \gg 1) \). We interpret \(|\lambda|\) as the time to reach the consistency when \( \lambda < 0 \). Then in what follows, the absolute value of \( \lambda \) is discussed, in addition to its sign in this paper.

### 3. Numerical Simulation

In this section, we study when and how consistency occurs in systems where the inputs \( \epsilon(t) \) generated by Eq. (2) with Eq. (3). In what follows, we study the two cases: chaotic and non-chaotic in the noise-free dynamics. First, as chaotic systems, we consider systems based on the Lorenz system and Rössler's second system. These two systems exhibit consistency in systems forced by the Ornstein-Uhlenbeck noises. Second, as a non-chaotic system, we consider a system based on the van der Pol oscillator.

#### 3.1. Lorenz System

The forced Lorenz system which we study in this subsection is described as follows.

\[
\begin{align*}
\dot{x} &= a(y - x), \\
\dot{y} &= bx - y - xz + \epsilon(t), \\
\dot{z} &= -cz + xy,
\end{align*}
\]

where \( a, b \) and \( c \) are real parameters. For the noise-free system \( (\epsilon(t) \equiv 0) \) with \( a = 10, b = 28 \) and \( c = 8/3 \), the largest Lyapunov exponent is known to be about 0.9. In this noise-free system, it has been known that (i) the typical frequency of \( \epsilon(t) \) is \( \omega := 2\pi f \sim 8.2 \), and that (ii) there are no typical frequencies for \( x(t) \) and \( y(t) \). In addition, for systems forced by noise, it has been shown to exhibit consistency when \( \epsilon(t) \) is either Gaussian white noise [5] or the Ornstein-Uhlenbeck noise [6].
Here, in the system forced by $\varepsilon(t)$ generated by Eq.(2) with Eq. (3), we show that the system exhibits consistency. In Fig. 1(a), the consistency is observed whenever $\sigma$ is greater than $\sigma_0$. In the regime $\gamma \gg 1$, we observe that $\sigma_c$ becomes larger as $\gamma$ increases. The reason why this increase of $\sigma_c$ with $\gamma$ appears can be ascribed to the nature of the input-noise, as explained in Ref. [6], where they studied the case that the Ornstein-Uhlenbeck noise is applied to systems.

In Fig. 1(b), the contour plot of the largest conditional Lyapunov exponent in the $\gamma \omega_0-\gamma$ plane is shown. It is clear that the regime where $\lambda$ is small is located near a certain value of $\gamma \omega_0$. We denote such value of $\gamma \omega_0$ by $\omega_\lambda$ (see the denominator of Eq.(5)). In this case, we find $\omega_\lambda \sim 8.2$. It should be noted here that this value, $\omega_\lambda \sim 8.2$, is the same as the typical frequency of the noise-free Lorenz system. Accordingly, the system is in resonance with the input signal generated by Eq.(2). In addition, from Eq.(6), we cannot expect that this resonance phenomenon occurs when the Ornstein-Uhlenbeck signal is applied to the system, because only $\omega = 0$ gives the maximal power spectrum of the noise.

In addition to the present example, we now illustrate that such resonance phenomenon occurs in some cases. First we consider the system with $a = 10$, $b = 28$ and $c = 0.8$. In the noise-free system, the typical frequency is about 4.5. For the system forced by $\varepsilon(t)$ generated by Eq.(2) with Eq.(3), we confirm that both $\sigma_c$ becomes larger as $\gamma$ increases, and the resonance phenomenon is observed around $\omega_\lambda \sim 4.5$, as shown in Figs. 2(a) and (b) respectively. Furthermore, in the next subsection, we study another system to illustrate that there is the resonance phenomenon in addition to the forced Lorenz system.

3.2. Rössler’s second System

For the case where noise-free systems are chaotic, we would like to check that resonance phenomena occur not only in the forced Lorenz system but also another forced system in this subsection. The system which we study here is based on the Rössler’s second system[9], and our system is described as

$$
\begin{align*}
\beta^{-1} \dot{x} &= x - xy - z, \\
\beta^{-1} \dot{y} &= x^2 - ay + \varepsilon(t), \\
\beta^{-1} \dot{z} &= b(cx - z),
\end{align*}
$$

where $\beta, a, b$ and $c$ are real parameters. We consider the case $\beta = 10, a = 0.1, b = 0.08$ and $c = 0.125$. It is noted here that $\beta$ controls the time-scale of the system. For the noise-free system, the phase portrait is depicted in Fig. 3(a), and the system is shown to be chaotic. As shown in Fig.3(b), a typical frequency $\omega$ is about 3.8. For the system forced by the Ornstein-Uhlenbeck noise, the system exhibits consistency (no figure given). To discuss the resonance phenomenon, we numerically calculate $\lambda$ in the system forced by $\varepsilon(t)$ generated by Eq. (2) with Eq. (3).

In Figs. 3(c) and (d), the same tendencies, which we have already observed in §3.1, are again obtained here. That is, $\sigma_c$ becomes larger as $\gamma$ increases, shown in Fig.3(c), and the resonance phenomenon occurs, as shown in Fig.3(d). Then we expect that the resonance phenomenon occurs in a system exhibiting consistency, which is chaotic in the noise-free case.

In the next subsection, we study the other class: noise-free systems are non-chaotic.

3.3. Van der Pol oscillator

We are now interested in whether or not we observe the resonance phenomenon even for the case that the noise-free systems are non-chaotic.
The system which we study here is defined as

\[ \ddot{x} + (x^2 - 1) \dot{x} + x = \varepsilon(t), \]

where \( \varepsilon(t) \) represents an input signal. When the system is noise-free, the system is called the van der Pol oscillator that has a limit cycle. For the noise-free system, the typical time sequence of \( x(t) \) is shown in Fig. 4(a), and the frequency is \( \omega = 2\pi f \sim 0.93. \) For the system forced by the Ornstein-Uhlenbeck noise, it has been shown that the largest conditional Lyapunov exponent is always negative[6].

Here we concentrate on the case where \( \varepsilon(t) \) is generated by Eq.(2) with Eq. (3). The negativity of \( \lambda \) in Fig. 4(b) is consistent with one of results in Ref.[7]. Contrary to the case that a noise-free system is chaotic, we observe no resonance phenomenon. Our results indicates that the resonance phenomenon is only observed in systems exhibiting consistency, which are chaotic in the noise-free case.

Due to the studies of two classes, chaotic and non-chaotic systems, we have the following conjecture: for any chaotic system, which exhibits consistency, the resonance phenomenon occurs.

4. Conclusion

We have studied whether or not consistency occurs in systems driven by a common additive colored noise. The colored noise which we use in this paper is generated by a linear second-order differential forced by Gaussian white noise. For the case that noise-free systems are chaotic, we have found that there is a resonance phenomenon, but not in the case that a noise-free system is non-chaotic. In addition, it has been numerically shown that the resonance frequency is the frequency of the noise-free system under study.

Because the input to a system could be expressed in the frequency space, we could expect that a similar resonance phenomenon occurs, when we study a system forced by a general input.

Although it remains questions when and how consistency occurs in much more general situations, this study might give a clue to understand the complex behavior of consistency. For this reason, we believe that this study leads to a more complete understanding of the nature of consistency.

References

Numerical Verification of Five Solutions in Two-transistor Circuits

Yusuke Nakaya†, Shin’ichi Oishi†, Tetsuo Nishi† and Martin Claus‡

†Faculty of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
‡Department of Electrical Engineering and Information Technology, Dresden University of Technology
D-01062 Dresden, Germany
Email: nakaya@waseda.jp, oishi@waseda.jp, nishi-t@waseda.jp, claus@iee.et.tu-dresden.de

Abstract—This paper is concerned with the number of operating points in two-transistor circuits. Recently, a new method was proposed for analyzing two-transistor circuit and an Ebers-Moll BJT circuit possessing five operating points was shown numerically. In this paper, applying the technique of numerical validation, we verify the numerical results rigorously and prove that the circuit have more than three operating points rigorously.

1. Introduction

One of the most fundamental problems on resistive transistor circuits is to find the maximum number of solutions of $m$-transistor circuits and it has long been studied [1]–[6] since the famous Nielsen-Willson theorem on a unique solution [1]–[9]. There are some conjectures on the maximum number of solutions. For example, it may be $2^m - 1$, etc. It was once proposed as “a Challenging Problem” in the field of Nonlinear Circuits and Systems, but it still remains unsolved even for the simplest case of $m = 2$.

The number of solutions of course depends on various factors, for example, types of transistors (BJT or MOST), modeling of a transistor (Ebers-Moll or Gummel-Poon), performance of circuit simulators, the characteristics of a diode function $f$ (exponential function, piecewise-linear function, etc) in transistor models. The difficulties of the problem can be seen from [10]–[12].

One of the most well-known results so far, which has long been believed to be true, was that the maximum number of solutions of two-transistor circuits is three. The proof by Lee [2]–[3] is extremely complex to fully understand. Afterwards Goldgeisser et al [7] showed that Lee’s theorem is not logically true for MOST circuits and then constructed a five-solution MOST circuit (a modified latch circuit having the feedback structure). After that Shou et al [8] constructed even for BJT circuits a five-solution circuit by using circuit simulator model. Recently Claus [9] proposed a new analysis method for a two-transistor circuit which is based on the notion of terminal behavior of a network. Applying this method on a MOST circuit and an Ebers-Moll BJT circuit he showed numerically that these circuits possess a least five solutions. Furthermore, the results are verified with homotopy methods [9] realizable with standard circuit simulators.

As stated above, all five-solution circuits so far found are derived based on theoretically predicted observations and are numerically examined. However in general the numerical computation always involves some numerical errors and the accuracy of computation depends on the ability of circuit simulators, etc. Indeed the numerical results by Spice and Matlab differs considerably (private communication). So if we intend to confirm these results surely, the numerical results have to be verified carefully.

Though the numerical validation is usually considered to require much computation time, we are recently developing various effective tools for fast numerical validation. So the aim of this paper is to give a methodology for validating the numerical results rigorously and to show that, as an example, the Claus’ results can be rigorously proved to be true by numerical analysis.

2. Fundamental Technique for the Numerical Verification

Let us consider about numerical verification of a solution for a system of nonlinear equations

$$f(x) = 0, \quad (1)$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$. For the purpose, we can apply Krawczyk’s method which is known as one of the efficient method to prove the existence of the solution [14]. In this section, this Krawczyk’s method is introduced for the verification of solutions in the following section.

For an interval $T$ whose center is $c$, an interval ma-
trix $M$ and an interval mapping $K$ are defined as
\begin{equation}
M = E - L^{-1}F'(T)
\end{equation}
and
\begin{equation}
K(T) = c - L^{-1}f(c) + M(T - c)
\end{equation}
respectively, where $E$ is an $n \times n$ unit matrix, $L$ is an approximate matrix of $f'(c)$, and $F'$ is an interval inclusion of $f'$. Then, if the condition
\begin{equation}
K(T) \subset T
\end{equation}
holds, there exists a unique solution $x^*$ of the nonlinear equations (1) in the interval $T$.

Here, let $c$ be an approximate solution of (1) which is obtained by an appropriate method. Define $T$ as
\begin{equation}
T = c + X
\end{equation}
where, $X$ is described as
\begin{equation}
X = [-c, c],
\end{equation}
and $e$ is defined as
\begin{equation}
e = 2\|L^{-1}f(c)\|_{\infty}.
\end{equation}
Using the Krawczyk operator (3), define the operator $H$ as
\begin{align*}
H &= K(T) - c \\
&= K(c + X) - c \\
&= -L^{-1}f(c) + (E - L^{-1}F'(c + X))X.
\end{align*}
At this time, the condition (4) is equivalent to
\begin{equation}
H \subset X.
\end{equation}
Therefore, in the following sections, we apply the operator $H$ and check the condition (9) for proving the existence of solution.

3. Analysis of Two-transistor Circuit

In [9], Claus has analyzed a two-transistor circuit shown in figure 1, and it is reported that this circuit has five operating points under certain conditions. In this section, we analyze this two-transistor circuit, and attempt to prove that this circuit has certainly five operating points by calculating with error verification.

For calculating operating points of the circuit numerically, we apply the Ebers-Moll model as a model of transistor. A transistor is represented as figure 2 by Ebers-Moll model, and terminal current is given as
\begin{equation}
\begin{bmatrix}
-I_1 \\
-I_2
\end{bmatrix} = \begin{bmatrix}
1 & -\alpha_r \\
-\alpha_f & 1
\end{bmatrix} \begin{bmatrix}
f_1(V_1) \\
f_2(V_2)
\end{bmatrix},
\end{equation}
where
\begin{align*}
\begin{pmatrix}
f_1(V_1) \\
f_2(V_2)
\end{pmatrix} &= \begin{bmatrix}
\frac{I_c}{\alpha_f} \left( e^{\frac{V_1}{V_T}} - 1 \right) \\
\frac{I_c}{\alpha_r} \left( e^{\frac{V_2}{V_T}} - 1 \right)
\end{pmatrix}.
\end{align*}
Applying (10), the equation of the circuit shown in figure 1 is given as a system of nonlinear equations as follows:
\begin{equation}
Tf(V) + GV + J = 0,
\end{equation}
where
\begin{equation}
T = \begin{bmatrix}
1 & -\alpha_r & 0 & 0 \\
-\alpha_f & 1 & 0 & 0 \\
0 & 0 & 1 & -\alpha_r \\
0 & 0 & -\alpha_f & 1
\end{bmatrix},
\end{equation}
and
\begin{equation}
f(V) = \begin{bmatrix}
\frac{I_c}{\alpha_f} \left( e^{\frac{V_1}{V_T}} - 1 \right) \\
\frac{I_c}{\alpha_r} \left( e^{\frac{V_2}{V_T}} - 1 \right) \\
\frac{I_c}{\alpha_f} \left( e^{\frac{V_3}{V_T}} - 1 \right) \\
\frac{I_c}{\alpha_r} \left( e^{\frac{V_4}{V_T}} - 1 \right)
\end{bmatrix}.
\end{equation}
\[ G = \begin{bmatrix} 2G_b + G_c & -(G_b + G_c) & -2G_b & G_b \\ -(G_b + G_c) & G_b + G_c & G_b & 0 \\ -2G_b & G_b & 2G_b + G_c & -(G_b + G_c) \\ G_b & 0 & -(G_b + G_c) & G_b + G_c \end{bmatrix}, \] (15)

\[ V = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \\ V_4 \end{bmatrix}, \] (16)

\[ J = \begin{bmatrix} G_c V_c \\ G_s V_s - G_c V_c \\ G_s V_s - G_c V_c \end{bmatrix}, \] (17)

and \( G_b, G_c \) represent \( \frac{1}{R_b}, \frac{1}{R_c} \) respectively.

We analyze the case that the values of \( R_b, R_c \) and \( V_{cc} \) are 10kΩ, 5kΩ and -5V respectively and \( V_s \) works as a parameter. On the other hand, for the parameters of transistor, we choose \( \alpha_f = 0.99, \alpha_r = 0.5 \), and we consider following two cases: (a) \( I_s = 10^{-9}(A), V_T = 0.053(V) \), (b) \( I_s = 10^{-6}(A), V_T = 0.102(V) \).

On the first step, we calculate the solutions of the equation (12) by normal numerical calculation, namely without verifying errors. By solving the equation (12) for different values of \( V_s \), we obtain following bifurcation diagrams. Figure 3 shows the case (a), and Figure 4 shows the case (b). On these bifurcation diagrams, the number of operating points is indicated

![Bifurcation diagrams for the case (a)](image1)

![Bifurcation diagrams for the case (b)](image2)

Figure 4: Bifurcation diagrams for the case (b)

Figure 3: Bifurcation diagrams for the case (a)

as the number of intersection points of the bifurcation diagram with a vertical line corresponding to a specific value of \( V_s \). It is obviously recognized that there exists an interval for \( V_s \) in which there are five operating points for both cases.

From these results, on the next step, we prove rigorously that the two-transistor circuit shown in figure 1 have at least five operating points. From figure 3 and 4, we consider the case \( V_s = -0.64(V) \) for case (a) and \( V_s = -0.44(V) \) for case (b). For the purpose, we apply Krawczyk’s method to the nonlinear equation (12). As approximate solutions for equation (12), we choose intersection points of the bifurcation diagram and vertical line of \( V_s \). Applying Krawczyk’s method, five solutions are calculated rigorously for each cases of (a) and (b). These results are shown in table 1 and 2. From this results, it is rigorously proved that the two-transistor circuit shown in figure 1 have at least five operating points.

4. Conclusion

In this paper, we were concerned with the number of operating points in two-transistor circuits. For proving that there are some two-transistor circuits which have more than three operating points, an Ebers-Moll BJT circuit was analyzed. Applying the technique of
Table 1: Verified solutions of case (a) on \( V_s = -0.64(V) \)

<table>
<thead>
<tr>
<th>No.</th>
<th>( V_1 )</th>
<th>( V_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>[0.74231646, 0.74231648]</td>
<td>[0.61988729, 0.61988732]</td>
</tr>
<tr>
<td>#2</td>
<td>[0.73551252, 0.73551255]</td>
<td>[0.57555294, 0.57555297]</td>
</tr>
<tr>
<td>#3</td>
<td>[0.72928957, 0.72928960]</td>
<td>[0.51630182, 0.51630185]</td>
</tr>
<tr>
<td>#4</td>
<td>[0.71990159, 0.71990162]</td>
<td>[-0.00333576, -0.00333579]</td>
</tr>
<tr>
<td>#5</td>
<td>[0.70358952, 0.70358955]</td>
<td>[-1.36664107, -1.36664110]</td>
</tr>
</tbody>
</table>

Table 2: Verified solutions of case (b) on \( V_s = -0.44(V) \)

<table>
<thead>
<tr>
<th>No.</th>
<th>( V_1 )</th>
<th>( V_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>[0.72888014, 0.72888016]</td>
<td>[0.51630182, 0.51630185]</td>
</tr>
<tr>
<td>#2</td>
<td>[0.70816927, 0.70816930]</td>
<td>[0.40346502, 0.40346505]</td>
</tr>
<tr>
<td>#3</td>
<td>[0.69521428, 0.69521431]</td>
<td>[0.24620350, 0.24620353]</td>
</tr>
<tr>
<td>#4</td>
<td>[0.67667878, 0.67667900]</td>
<td>[-0.20824492, -0.20824495]</td>
</tr>
<tr>
<td>#5</td>
<td>[0.61902389, 0.61902392]</td>
<td>[-1.36664114, -1.36664117]</td>
</tr>
</tbody>
</table>

Numerical validation, we verified numerically that the circuit have at least five operating points rigorously.

References


Error bounds for extremely ill-conditioned problems

Siegfried M. Rump

Institute for Reliable Computing, Hamburg University of Technology,
Schwarzenbergstraße 95, Hamburg 21071, Germany
and Waseda University, Faculty of Science and Engineering,
2-4-12 Okubo, Shinjuku-ku, Tokyo 169-0072, Japan.
email: rump@tu-harburg.de.

Abstract— We discuss methods to compute error bounds for extremely ill-conditioned problems. As a model problem we treat matrix inversion. We demonstrate that additive corrections to improve an approximate inverse are useful for ill-conditioned problems, but hardly usable for extremely ill-conditioned problems. Here multiplicative corrections can be used, including the possibility to compute guaranteed error bounds.

1. Introduction

Let us consider the inversion of a real $n \times n$ matrix as a model problem. This can be interpreted as a function from $\mathbb{R}^{n^2}$ to $\mathbb{R}^{n^2}$, which is continuous and differentiable on the set of $n^2$-vectors which, interpreted as an $n \times n$ matrix $A$, are nonsingular. Suppose $A$ is nonsingular. The problem of matrix inversion is called ill-conditioned when small changes in the input data $A$ cause large changes in the solution. If $A$ is a matrix of floating point numbers and changes in the unit of the last place of the input data $A$ cause a relative change in the solution of the order of 100%, then the problem is extremely ill-conditioned. It is common to believe that such problems cannot be solved in the same floating point format as the input data is stored. This fact is supported by the well known relation

$$\text{forward error} = \text{condition} \times \text{backward error}.$$  \hfill (1)

Suppose the floating point format in use is IEEE 754 double precision corresponding to 53 bits in the mantissa. Then this means that the solution of problems with condition number $2^{53} \approx 10^{16}$ or above cannot be reasonably approximated in double precision. However, there are exceptions to that statement, namely, if, for some reason, intermediate operations can be performed exactly without rounding error, for instance in integer computations. A very interesting example are so-called error-free transformations. Here pairs of floating point numbers are transformed into pairs of floating point numbers without error. Consider the following algorithm by Knuth, 1969 [2]:

$$\text{function } [x, y] = \text{TwoSum}(a, b)$$
$$x = \text{fl}(a + b)$$
$$z = \text{fl}(x - a)$$
$$y = \text{fl}((a - (x - z)) + (b - z))$$

This algorithm satisfies the following property.

**Theorem 1** Given two floating point numbers $a, b \in \mathbb{R}$, the result $(x, y)$ of TwoSum satisfies

$$x = \text{fl}(a + b) \quad \text{and} \quad x + y = a + b.$$  

Note that this statement is also true in the presence of underflow. Using TwoSum we can present the following algorithm for vector transformation.

$$\text{function } p = \text{VecTransform}(p)$$
$$\text{for } i = 2 : n \quad \% \text{length}(p) = n$$
$$[p_i, p_{i-1}] = \text{TwoSum}(p_i, p_{i-1})$$

Note that we use the same variable $p$ for input and output. One sees immediately by induction that the value of the sum $s := \sum p_i$ is not changed by VecTransform. For such an algorithm the relation (1) cannot hold since all intermediate transformations are exact. One can show [3] that the condition number of $\sum(p)$ drops basically by a factor $2^{-53}$ with each application of VecTransform. After sufficiently many transformations, $\text{fl}(\sum p)$ will be a very accurate approximation.
of the exact sum $s$. Such algorithms are presented in [3] and [6].

The following ingenious method by Dekker splits a 53-bit floating point number $a \in \mathbb{F}$ into two 26-bit parts.

function $[x, y] = \text{Split}(a)$
\[
    c = \text{fl}(\text{factor} \cdot a) \quad \%\text{factor} = 2^{27} + 1
    
    x = \text{fl}(c - (c - a))
    
    y = \text{fl}(a - x)
\]

The result of this algorithm satisfies $x + y = a$ such that the results $x, y$ do have not more that 26 significant bits. The trick is that the sign bit of $x$ or $y$ is used for the representation. The function Split is used to transform the product of two floating point numbers into the sum of two floating point numbers, again an error-free transformation. This is done by the following algorithm TwoProduct.

function $[x, y] = \text{TwoProduct}(a, b)$
\[
    x = \text{fl}(a \cdot b) \\
    [a_1, a_2] = \text{Split}(a) \\
    [b_1, b_2] = \text{Split}(b) \\
    y = \text{fl}(a_2 \cdot b_2 - (((x - a_1 \cdot b_1) - a_2 \cdot b_1) - a_1 \cdot b_2))
\]

The mathematical property
\[
x + y = a \cdot b
\]

is true for all $a, b \in \mathbb{F}$ as long as no over- or underflow occurs. With this the dot product of two n-vectors can be transformed into the sum of an 2n-vector, which in turn can be summed up with the methods described.

Therefore we can calculate accurate approximations of the $\sum(p_i)$ and $x^T y$ for n-vectors $p, x, y$. Note that the mentioned algorithms use only ordinary floating point addition and multiplication. The algorithms proved to be very fast [3, 6].

Let a matrix $A \in \mathbb{F}^{n \times n}$ be given. One may ask whether it is possible to compute an accurate approximation of the inverse of $A$ using accurate summation and dot product.

2. Additive corrections

Let $R$ be an approximate inverse of $A$, for example computed by the Matlab command $\text{inv}(A)$. The obvious choice of iterative improvement of $R$ is a Newton iteration. This method is known as Schulz iteration in the literature [7]. The first step of the iteration as follows.

\[
\begin{align*}
    X & := R \\
    X & := X - (X \cdot A - I) \cdot X
\end{align*}
\]

(2)

Here $I$ denotes the $n \times n$ identity matrix. Consider as a model problem a randomly generated $100 \times 100$ matrix of condition number $10^{10}$. We used \texttt{randsvd} as described in [1, Chapter 28]. In double precision with relative rounding error unit $u = 2^{-53}$ we can expect $\log_2(u^{-1}/10^{10}) \sim 6$ correct digits of the solution. We check the accuracy by means of the residual $\|I - R \cdot A\|$. We also check the right residual and obtain

\[
\begin{align*}
    \|I - R \cdot A\| & = 9.93 \cdot 10^{-7} \quad \text{and} \\
    \|I - A \cdot R\| & = 7.87 \cdot 10^{-6}
\end{align*}
\]

(3)

We display the result for one typical example. Note that we used high precision to calculate the residuals, so we can expect them to be correct. As expected the left residual is slightly better since the Matlab routine $\text{inv}$ computes a left inverse of $A$. Next we perform one iteration (2) in double precision. We obtain

\[
\begin{align*}
    \|I - X \cdot A\| & = 6.30 \cdot 10^{-7} \quad \text{and} \\
    \|I - A \cdot X\| & = 1.54 \cdot 10^{-2}
\end{align*}
\]

(4)

First, the residual $I - X \cdot A$ in (2) is computed in double precision. This computation is always ill-conditioned, also for well-conditioned matrices, so we cannot expect a significant improvement. The left residual improves slightly, whereas the quality of $X$ as a right inverse deteriorates completely. This is because the residual in (2) is written for a left inverse. The corresponding iteration for the right inverse

\[
\begin{align*}
    X & := R \\
    X & := X - X \cdot (A \cdot X - I)
\end{align*}
\]

(5)

yields

\[
\begin{align*}
    \|I - X \cdot A\| & = 1.40 \cdot 10^{-2} \quad \text{and} \\
    \|I - A \cdot X\| & = 8.06 \cdot 10^{-7}
\end{align*}
\]

(6)

as expected. Next we concentrate on the left inverse and compute the residual $I - X \cdot A$ in quadruple precision. By a common rule of thumb we may expect an improvement of the relative accuracy of the result by $\text{cond} \cdot u \sim 10^{-6}$. We denote this by

\[
\begin{align*}
    X & := R \\
    X & := X - \text{double}(\text{quad}(X \cdot A - I)) \cdot X
\end{align*}
\]

(7)
Here \( \text{quad}(\cdot) \) indicates that the whole expression inside the parantheses is computed in quadruple precision, and \( \text{double}(\cdot) \) means to round it back into double precision. The result is as follows.

\[
\begin{align*}
\| I - X \cdot A \| &= 1.64 \cdot 10^{-7} \quad \text{and} \\
\| I - A \cdot X \| &= 2.06 \cdot 10^{-7}.
\end{align*}
\]  

(8)

Interestingly, now the right residual is of the same quality as the right residual, however, no improvement is visible. The next step is to perform the multiplication by \( X \) in quadruple as well.

\[
\begin{align*}
X &:= R \\
X &:= X - \text{double}(\text{quad}(X \cdot A - I) \cdot X).
\end{align*}
\]  

(9)

The result is as follows.

\[
\begin{align*}
\| I - X \cdot A \| &= 2.41 \cdot 10^{-7} \quad \text{and} \\
\| I - A \cdot X \| &= 3.09 \cdot 10^{-7}.
\end{align*}
\]  

(10)

Again we see no improvement. Finally, we may execute the whole computation of \( X \) in (2) in quadruple precision and round the result to double precision:

\[
\begin{align*}
X &:= R \\
X &:= \text{double}(\text{quad}(X - (X \cdot A - I) \cdot X)).
\end{align*}
\]  

(11)

The result is as follows.

\[
\begin{align*}
\| I - X \cdot A \| &= 2.09 \cdot 10^{-7} \quad \text{and} \\
\| I - A \cdot X \| &= 2.96 \cdot 10^{-7}.
\end{align*}
\]  

(12)

Strange enough the Newton iteration does not improve the result at all although all computations have been performed in quadruple precision and only the final result is rounded into double precision. Finally, we use the nearest floating point matrix to the exact inverse \( A^{-1} \), the latter computed by some multiple precision routine, that is

\[
X := \text{double}(A^{-1}).
\]  

(13)

Then the left and right residuals are

\[
\begin{align*}
\| I - X \cdot A \| &= 2.26 \cdot 10^{-7} \quad \text{and} \\
\| I - A \cdot X \| &= 3.60 \cdot 10^{-7}.
\end{align*}
\]  

(14)

So there seems not much chance to decrease the left or right residual using a double precision matrix \( X \) as preconditioner. However, a residual iteration using \( R \) needs the residual matrix to be convergent. For extremely ill-conditioned matrices a double precision \( X \) seems unsufficient. A remedy is to store \( X \) in two parts. This was already used in [5] and was later called staggered correction. One method is to store the original \( X \) and the correction into two different parts.

\[
\begin{align*}
X &:= R \\
X_1 &:= X, \\
X_2 &:= \text{double}(\text{quad}(X \cdot A - I) \cdot X).
\end{align*}
\]  

(15)

Note that only the residual is computed in quadruple precision and rounded into double, the multiplication by \( X \) is in double precision. Now the result is as follows.

\[
\begin{align*}
\| I - (X_1 + X_2) \cdot A \| &= 1.18 \cdot 10^{-13} \quad \text{and} \\
\| I - A \cdot (X_1 + X_2) \| &= 1.17 \cdot 10^{-12}.
\end{align*}
\]  

(16)

As a result we see that for ill-conditioned problems the preconditioner \( R \) still contains enough information to produce small residuals, but only if the new preconditioner is stored in multiple precision. For extremely ill conditioned problems, however, things change again. Let \( A \) be a \( 100 \times 100 \) matrix with condition number \( 10^{15} \). This is at the limit of what can be solved in double precision. Then (15) with \( R = \text{inv}(A) \) yields

\[
\begin{align*}
\| I - (X_1 + X_2) \cdot A \| &= 4.09 \cdot 10^{+4} \quad \text{and} \\
\| I - A \cdot (X_1 + X_2) \| &= 8.87 \cdot 10^{+5},
\end{align*}
\]  

(17)

so no information at all. Also with the exact inverse by (13) all information is gone:

\[
\begin{align*}
\| I - X \cdot A \| &= 3.87 \cdot 10^{+4} \quad \text{and} \\
\| I - A \cdot X \| &= 4.27 \cdot 10^{+5}.
\end{align*}
\]  

(18)

The reason is that the initial approximation \( X \), an approximate inverse of \( A \), is so far from the true result \( A^{-1} \) that an additive correction does not work.

### 3. Multiplicative corrections

For a multiplicative correction we use the fact that an approximate inverse still contains a lot of structure. It does not contain adequate information to solve the residual equation directly, however, it may serve as a preconditioner. For the last example with the \( 100 \times 100 \) matrix \( A \) with condition number \( 10^{15} \) we use multiple precision arithmetic to compute \( \text{cond}(A) \), \( X := R \cdot A \).
and \( \text{cond} \,(X \cdot A) \), where \( R \) is the approximate inverse if \( A \) by \( R := \text{inv}(A) \).

\[
\begin{align*}
\text{cond} \,(A) &= 3.92 \cdot 10^{21} \\
\text{cond} \,(X \cdot A) &= 2.69 \cdot 10^{7}.
\end{align*}
\]  

It is a general observation that for an arbitrarily ill-conditioned matrix \( A \) we can compute an approximate inverse \( R := \text{inv}(A) \) in double precision such that the condition number of the preconditioned matrix \( X := R \cdot A \) drops by roughly a factor \( u \). The product \( R \cdot A \) is computed using algorithms presented in [3] or [6]. This corresponds to a result as if computed in quadruple precision and then rounded to double precision or, with faithful rounding, respectively. The dropping of the condition number is also observed when computing \( R \cdot A \) only in double precision.

These observations were used in about 1984, when we derived a method for inverting arbitrarily ill-conditioned matrices. This method, which we never published, requires the possibility to calculate a dot product \( x^T y \) in \( k \)-fold precision and store into working precision as in [3] or [6]. The first step of the method is as follows.

\[
\text{function \( [X1, X2] = \text{AccInv}(A) \))}
\]

\[
R = \text{inv}(A) \quad \% \text{double precision}
\]

\[
P = \text{inv(double}(\text{quad}(R \cdot A)))
\]

\[
X1 + X2 = \text{quad}(P \cdot R)
\]

As we have seen before we need higher precision for a preconditioner to produce a small residual. The last line in our algorithm computes the product \( P \cdot R \) in quadruple precision and stores the result in two double precision parts \( X1, X2 \). This produces

\[
\begin{align*}
\| I - (X1 + X2) \cdot A \| &= 3.17 \cdot 10^{-9} \quad \text{and} \\
\| I - A \cdot (X1 + X2) \| &= 2.03 \cdot 10^{+7}.
\end{align*}
\]  

The left residual is small enough to produce a converge iteration, for example, for solving a system of linear equations. The right residual is large since \( P \) is computed using \( R \) as a left preconditioner. Calculating \( P \) as an approximate inverse of \( A \cdot R \) and \( X \) by \( R \cdot P \) yields

\[
\begin{align*}
\| I - (X1 + X2) \cdot A \| &= 2.02 \cdot 10^{+8} \quad \text{and} \\
\| I - A \cdot (X1 + X2) \| &= 3.56 \cdot 10^{-5},
\end{align*}
\]  

as expected. The presented algorithm AccInv is the first step of an iterated algorithm, the latter being able to compute an approximate inverse of an arbitrarily ill-conditioned matrix in double precision floating point. It uses only the basic operations in double and an accurate dot product as presented in [3] or [6]. Very recently a partial analysis of this algorithm was presented in [4]. It is still quite mysterious how purely double precision computation can invert arbitrarily ill-conditioned matrices, and a full analysis is still open.

References


Convergence Theorem of Rump’s Method for Inverting Arbitrarily Ill-Conditioned Matrices

Shin’ichi Oishi †, Kunio Tanabe ‡, Takeshi Ogita †∗ and Siegfried M. Rump †‡

† CREST, Japan Science and Technology Agency
‡ Faculty of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
‡ Institute for Reliable Computing, Hamburg University of Technology,
Schwarzenbergstr. 95, Hamburg 21071, Germany
Email: |oishi, tanabe.kunio, ogita}@waseda.jp, rump@tu-harburg.de

Abstract—In this paper, the problem of inverting regular matrices with arbitrarily large condition number is treated in double precision defined by IEEE 754 floating point standard. In about 1984, Rump derived a method for inverting arbitrarily ill-conditioned matrices. Numerical experiments exhibit that Rump’s method converges rapidly for various matrices with large condition numbers. Why Rump’s method is so efficient for inverting arbitrarily ill-conditioned matrices is a little mysterious. Thus, to prove its convergency is an interesting problem in numerical error analysis. In this article, we shall present a convergence theorem for a variant of Rump’s method. Numerical experiments are presented for illustrating the validity of our numerical error analyses.

2. Convergence Theorem

We assume that the dimension of the problem, \( n \), satisfies \( nu \ll 1 \) and \( C_i \sqrt{u} \ll 1 \). In this paper, \( C_i, \ i = 0, 1, 2, \ldots \) denote numbers of \( O(1) \) satisfying \( C_iu \ll 1 \) and \( C_i \sqrt{u} \ll 1 \). Moreover, \( c_n \) denotes a numbers of \( O(n) \) satisfying \( c_nu \ll 1 \) and \( c_n \sqrt{u} \ll 1 \).

Let \( A = (a_{ij}) \) be a real \( n \times n \) matrix and \( \Pi = (\pi_{ij}) \) be an approximate inverse of \( A \). Let \( b \in \mathbb{R}^n \) and \( \tilde{x} \) be an approximate solution of \( Ax = b \). It is known that if

\[
\|\Pi A - I\| < 1
\]

is satisfied, \( A \) becomes regular. Here, \( I \) is the \( n \times n \) identity matrix and \( \| \cdot \| \) is a subordinate matrix norm. Rump’s method is an algorithm to produce \( \Pi \). Thus, from the above mentioned fact, we set a purpose of this paper to show that \( \Pi \) generated by Rump’s method eventually satisfies (1).

For the purpose, we introduce an accurate dot product calculation algorithm. Let \( A, B \in \mathbb{R}^{n \times n} \). Let us assume that we have an accurate dot product algorithm which calculates \( D_i \in \mathbb{R}^{n \times n}, i = 1, 2, \ldots, k \), satisfying

\[
\left| \sum_{i=1}^{k} D_i - AB \right| \leq C_0 u^k |AB|.
\]

Here, \( AB \) is the usual (error free) matrix multiplication and \( C_0 \) is a constant satisfying \( C_0 = O(1) \). We denote such an algorithm as

\[
D_{1:k} = [AB]_k \quad \text{with} \quad D_{1:k} = D_1 + D_2 + \ldots + D_k, \quad D_i \in \mathbb{R}^{n \times n}.
\]

A very efficient method for calculating such a dot product in \( k \)-fold accuracy was just developed in [11]. It uses only floating point operations in working precision, has no branches and is very fast.
In this paper, to simplify the life, working precision is assumed to be the double precision defined by IEEE 754 floating point standard. In the following, we use a variant of Rump’s method as given by the following Algorithm 1, which is written in Matlab-like:

**Algorithm 1 Modified Rump’s Method I**

\[ S_0 = A + ΔA; \]
\[ X_0 = \text{inv}(S_0); \quad Π_1 = X_0; \]
for \( k = 1 : k_{\text{max}} \)
\[ C = [Π_{1:k}A]; \]
\[ \tilde{S}_k = C + ΔC; \]
\[ X_k = \text{inv}(\tilde{S}_k); \]
\[ Π_{1:k+1} = [X_k Π_{1:k}]_{:+1}; \]
end

Here, \( \text{inv}(B) \) is a built-in function in Matlab for inverting \( B \in \mathbb{R}^{n \times n} \). Matrices \( ΔA \in \mathbb{R}^{n \times n} \) and \( ΔC \in \mathbb{R}^{n \times n} \) are defined by \( (ΔA)_{ij} = r_{ij} \sqrt{C_{ii}} \) and \( (ΔC)_{ij} = s_{ij} \sqrt{C_{ii}} \), respectively for all \( (i, j) \), where \( r_{ij} \) and \( s_{ij} \) are pseudo-random numbers distributed uniformly in \([-1, 1]\). Note that the perturbation \( ΔA \) and \( ΔC \) regularize \( A \) and \([Π_{1:k}A]\), respectively.

To simplify the notation, we will write \( Π_m \) instead of \( Π_{1:m} \) throughout the paper except in algorithms. We assume that all numerical calculation is done under IEEE 754’s double precision arithmetic in the nearest rounding mode.

Let \( S_k := Π_k A \). We now show that

\[ |S_k - \tilde{S}_k| \leq C_1 \sqrt{u}|S_k|, \tag{3} \]

where \( C_1 = O(1) \).

From (2), we have

\[ |S_k - [S_k]_1| \leq C_0 u|S_k|. \tag{4} \]

From the definition of \( ΔC \), it follows that

\[ |S_k - \tilde{S}_k| \leq |S_k - [S_k]_1| + |[S_k]_1 - \tilde{S}_k| \leq C_0 u|S_k| + \sqrt{u}|[S_k]_1|. \tag{5} \]

From

\[ |S_k| \leq |[S_k]_1| + |S_k - [S_k]_1| \leq |[S_k]_1| + C_0 u|S_k|, \tag{6} \]

we have

\[ |S_k| \leq \frac{1}{1 - C_0 u}|[S_k]_1|. \tag{7} \]

Moreover, from

\[ |[S_k]_1| \leq |\tilde{S}_k| + |[S_k]_1 - \tilde{S}_k| \leq |\tilde{S}_k| + \sqrt{u}|[S_k]_1|, \tag{8} \]

it follows that

\[ |[S_k]_1| \leq \frac{1}{1 - \sqrt{u}}|\tilde{S}_k|. \tag{9} \]

Substituting (7) and (9) into (5), it is seen that (3) holds with

\[ C_1 = \frac{1}{1 - \sqrt{u}}(1 + C_0 \sqrt{u}). \tag{10} \]

Using (3), we also have

\[ |S_k| \leq \frac{1}{1 - C_1 \sqrt{u}} |\tilde{S}_k|. \tag{11} \]

Since \( \tilde{S}_k \in \mathbb{R}^{n \times n} \), \( X_k \) can be computed by a standard inversion algorithm using Gaussian elimination in working precision.

The target of this section is to show that

\[ \kappa(S_{k+1}) = O(\sqrt{u})\kappa(S_k) + O(1) \tag{12} \]

provided that \( \kappa(S_k) \geq u^{-1} \).

For the purpose, in the first place, we estimate \( ||S_{k+1}||_\infty \) assuming that \( \kappa(S_k) \geq u^{-1} \). Let \( Γ := S_k - \tilde{S}_k \). Then, from (3) and (11) we have

\[ ||Γ||_\infty \leq C_1 \sqrt{u}||S_k||_\infty \leq C_1' \sqrt{u}||S_k||_\infty, \] \[ \text{where } C_1' := C_1/(1 - C_1 \sqrt{u}). \]

We note here that (13) states that the difference between \( S_k \) and \( \tilde{S}_k \), which is almost singular, is of order \( \sqrt{u}||S_k||_1 \). Thus, usually a distance between \( S_k \) and the nearest singularity, which lies very near to \( \tilde{S}_k \), becomes about \( C_1 \sqrt{u} \). This implies (cf. [3, 2])

\[ \kappa(\tilde{S}_k) = C_2 u^{-1/2}. \tag{14} \]

Here, we assume

**Assumption 1** \( C_2 = O(1) \).

This implies \( \kappa(\tilde{S}_k) = C_2 u^{-1/2} \ll u^{-1} \). Examples in the next section show that Assumption 1 is satisfied in many instances. Since a good approximate inverse of a matrix in \( \mathbb{R}^{n \times n} \) with a condition number much less than \( u^{-1} \) can be obtained in working precision, under Assumption 1 we can expect that \( X_k \) becomes a good approximate inverse of \( \tilde{S}_k \) satisfying

**Assumption 2** \( ||I - X_k \tilde{S}_k||_\infty = ε \ll 1 \).

We assume that Assumption 2 also holds. It follows from Assumption 2, \( \tilde{S}_k \) exists. Then, we note that

\[ ||X_k - S_k||_\infty = ||(I - X_k \tilde{S}_k)S_k^{-1}||_\infty \leq ||S_k^{-1}||_\infty ||I - X_k \tilde{S}_k||_\infty \leq \frac{||X_k||_\infty}{1 - ||I - X_k \tilde{S}_k||_\infty} ||I - X_k \tilde{S}_k||_\infty = \frac{ε}{1 - ε} ||X_k||_\infty. \tag{15} \]

From (15), it follows

\[ ||X_k||_\infty \leq ||S_k^{-1}||_\infty ||I - X_k - S_k^{-1}||_\infty \leq ||S_k^{-1}||_\infty + \frac{ε}{1 - ε} ||X_k||_\infty. \tag{16} \]

316
This and Assumption 2 imply that
\[
\|X_k\|_\infty \leq \frac{\|S_{k}^{-1}\|_\infty}{1 - \varepsilon} = \frac{1 - \varepsilon}{1 - 2\varepsilon}\|X_k\|_\infty = C_3\|S_{k}^{-1}\|_\infty. \tag{17}
\]
Here, \(C_3 := (1 - \varepsilon)/(1 - 2\varepsilon) = O(1)\). Let \(L\) and \(U\) be computed LU factors of \(S_k\). Then, since we have used Matlab’s ‘inv’ function, we have from [4, p. 268, (14.18)]
\[
\|I - X_k\tilde{S}_k\|_\infty \leq c_n u\|X_k\|_\infty\|L\|_\infty\|U\|_\infty, \tag{18}
\]
where \(c_n = O(n)\). Here, we introduce a constant \(g_k\) satisfying \(\|LU\|_\infty\|U\|_\infty \leq g_k\|S_k\|_\infty\). Then, we have
\[
\|I - X_k\tilde{S}_k\|_\infty \leq c_n g_k\|X_k\|_\infty\|\tilde{S}_k\|_\infty. \tag{19}
\]
From (14), (17) and (19), it follows that
\[
\|I - X_k\tilde{S}_k\|_\infty \leq c_n g_k C_3 u\|\tilde{S}_k\|_\infty = c_n C_4 \sqrt{u}, \tag{20}
\]
where \(C_4 := g_k C_3\). Under Assumption 2, which states \(\|I - X_k\tilde{S}_k\|_\infty \ll 1\), (20) asserts \(\|I - X_k\tilde{S}_k\|_\infty\) can be estimated as \(O(\sqrt{u})\) provided that \(C_4 = O(1)\). Thus, it turns out that Assumption 2 is equivalent to

**Assumption 3** \(C_4 = O(1)\) satisfying \(c_n C_4 \sqrt{u} \ll 1\).

Under this assumption, we show \(X_k\) is the exact inverse of \(\tilde{S}_k + \Delta\), where \(\|\Delta\|_\infty \leq c_n C_5 \sqrt{u}|S_k|_\infty\). Here, \(C_5\) is the constant defined below. From (20), we have for \(\Delta = X_k^{-1} - \tilde{S}_k\)
\[
\|\Delta\|_\infty = \|X_k^{-1} - \tilde{S}_k\|_\infty = \|X_k^{-1}(I - X_k\tilde{S}_k)\|_\infty \\
\leq \|X_k^{-1}\|_\infty\|I - X_k\tilde{S}_k\|_\infty \\
\leq \frac{\|\tilde{S}_k\|_\infty}{1 - \|I - X_k\tilde{S}_k\|_\infty}\|I - X_k\tilde{S}_k\|_\infty \\
\leq \frac{c_n g_k C_3 \sqrt{u}}{1 - c_n C_4 \sqrt{u}}\|\tilde{S}_k\|_\infty \leq c_n C_5 \sqrt{u}|S_k|_\infty. \tag{21}
\]
Here, using (13) we have put
\[
C_5 := \frac{C_4 C_3}{(1 - cn C_4 \sqrt{u})} = O(1). \tag{22}
\]

**Lemma 1** Let us assume that Assumptions 1 and 3 are satisfied. Then, the following a priori error estimate holds:
\[
\|I - X_k\tilde{S}_k\|_\infty \leq C_7, \tag{23}
\]
where \(C_7 := C_2 C_3 (C_1 + c_n g_k \sqrt{u})\). \tag{24}

**Proof.** Using (13), (17) and (19), we have
\[
\|I - X_k\tilde{S}_k\|_\infty = \|I - X_k(S_k - \tilde{S}_k + \tilde{S}_k)\|_\infty \\
\leq \|X_k(S_k - \tilde{S}_k)\|_\infty + \|I - X_k\tilde{S}_k\|_\infty \\
\leq C_1 \sqrt{u}\|X_k\|_\infty\|\tilde{S}_k\|_\infty + c_n g_k\|X_k\|_\infty\|S_k\|_\infty \\
\leq (C_1 + c_n g_k \sqrt{u}) \sqrt{u}\|X_k\|_\infty\|\tilde{S}_k\|_\infty \\
= C_6 \sqrt{u}\kappa(\tilde{S}_k). \tag{25}
\]
where \(C_6 := C_3 (C_1 + c_n g_k \sqrt{u})\). \tag{26}

This and (14) prove the lemma. \(\Box\)

From this lemma, we have
\[
\|X_k\tilde{S}_k\|_\infty \leq \|X_kS_k - f\|_\infty + \|f\|_\infty = 1 + \|I - X_k\tilde{S}_k\|_\infty \leq 1 + C_7. \tag{27}
\]

Here, we derive a relation between \(S_{k+1}\) and \(X_k\tilde{S}_k\):
\[
\|S_{k+1} - X_k\tilde{S}_k\|_\infty = \|\Pi_{k+1}A - X_k\Pi_k A\|_\infty = \|\Pi_{k+1} - X_k\Pi_k\|\|A\|_\infty \leq \|\Pi_{k+1} - X_k\Pi_k\|\|A\|_\infty \tag{28}
\]
Since \(\Pi_{k+1} = [X_k\Pi_k]_{k+1}\), we have
\[
\|\Pi_{k+1} - X_k\Pi_k\| \leq C_3 u^{k+1}\|X_k\Pi_k\|_\infty. \tag{29}
\]
Here, \(C_8 = O(1)\). Inserting this into (28), we have
\[
\|S_{k+1} - X_kS_k\|_\infty \leq C_3 u^{k+1}\|X_k\Pi_k\|\|A\|_\infty. \tag{30}
\]
Thus, we have
\[
\|S_{k+1}\|_\infty \leq \|X_k\tilde{S}_k\|_\infty + u^{k+1}\alpha, \tag{31}
\]
where \(\alpha := C_8\|X_k\|\|\Pi_k\|\|A\|_\infty\). \tag{32}

Here, we assume

**Assumption 4** \(u^{k+1}\alpha \ll 1\).

**Remark 1** Since \(\Pi_{k+1} = X_k\Pi_k\), usually we have
\[
\|\Pi_{k+1}\|_\infty \approx \|\Pi_k\|_\infty\|X_k\|_\infty. \tag{33}
\]
Here, \(\Pi_k, k = 1, 2, \ldots, \) work as the preconditioners for \(A\), we have \(\|S_k\|_\infty = \|\Pi_k A\|_\infty = O(1)\) and therefore \(\|S_{k}\|_\infty = O(1)\) for \(k \geq 1\). Thus, from (17)
\[
\|X_k\|_\infty \leq C_3 \|\tilde{S}_k\|_\infty = C_3 \kappa(\tilde{S}_k)\|\tilde{S}_k\|_\infty = O(u^{-1/2}) \tag{34}
\]
for \(k \geq 1\). Moreover, it can be expected that \(\tilde{S}_0\) is not so ill-conditioned and \(\kappa(\tilde{S}_0) = \|\tilde{S}_0\|_\infty\|\tilde{S}_0^{-1}\|_\infty = O(u^{-1/2})\), so that \(\|X_0\|_\infty = \|\text{inv}(\tilde{S}_0)\|_\infty \approx \|\tilde{S}_0^{-1}\|_\infty\). This and \(\|S_0\|_\infty \approx \|A\|_\infty\) yield
\[
\|X_0\|_\infty = O(u^{-1/2})\|A\|_\infty^{-1}. \tag{35}
\]
From (33), (34) and (35), it follows
\[
\|\Pi_k\|_\infty \approx O(u^{-1/2})\|A\|_\infty^{-1} \tag{36}
\]
provided that \(\kappa(S_k) \gg u^{-1}\). Thus, from (32) we have
\[
u^{i+1}\alpha \approx O(u^{i+1}u^{-1/2}(u^{1/2})^{-1})\|A\|_\infty^{-1}\|A\|_\infty = O(u^{i+1/2}). \tag{37}
\]
Under Assumption 4, it can be seen from (31) that
\[ \|S_{k+1}\|_\infty = \|X_k S_k\|_\infty + \epsilon, \quad (38) \]
where \( \epsilon \ll 1 \).

Now, we estimate \( \|S_{k+1}^{-1}\|_\infty \). Using (13), (21) and (22), we have
\[
\|(X_k S_k)^{-1}\|_\infty = \|(S_k + \Delta + \Gamma)^{-1} S_k^{-1}\|_\infty \\
\leq \|I + S_k^{-1}(\Delta + \Gamma)\|_\infty \\
\leq 1 + \|S_k^{-1}\|_\infty (\|\Delta\|_\infty + \|\Gamma\|_\infty) \\
\leq 1 + (C'_1 + c_k C_5) \sqrt{\mu}(\|S_k\|_\infty \|S_k^{-1}\|_\infty) \\
\leq 1 + (C'_1 + c_k C_5) \sqrt{\mu}(S_k). \quad (39)
\]

Let \( P \) and \( Q \) be regular \( n \times n \) matrices. If \( \|P - Q\|_\infty \leq \delta \), it follows that
\[ \|P^{-1} - Q^{-1}\|_\infty \leq \|P^{-1}(P - Q)Q^{-1}\|_\infty \leq \delta\|P^{-1}\|_\infty \|Q^{-1}\|_\infty. \quad (40) \]

Then, (30) and (40) yield
\[
\|S_{k+1}^{-1} - (X_k S_k)^{-1}\|_\infty \\
\leq \|S_{k+1}^{-1} - X_k S_k\|_\infty \|S_k^{-1}\|_\infty (X_k S_k)^{-1}\|_\infty \\
\leq u_k^{1/2} \|S_{k+1}^{-1}\|_\infty \|X_k S_k\|_\infty. \quad (41)
\]

where \( \beta := C_1 \|X_k\| \Pi + \|A\| \|S_k\|_\infty (X_k S_k)^{-1}\|_\infty \).

From (41), we have
\[
\|S_{k+1}^{-1}\|_\infty \\
\leq \|S_{k+1}^{-1} - (X_k S_k)^{-1}\|_\infty + \|X_k S_k\|_\infty (X_k S_k)^{-1}\|_\infty \\
\leq u_k^{1/2} \|S_{k+1}^{-1}\|_\infty + \|X_k S_k\|_\infty. \quad (42)
\]

If it holds that
**Assumption 5** \( u_k^{1/2} \beta \ll 1 \),
then we have
\[ \|S_{k+1}^{-1}\|_\infty \leq (1 - u_k^{1/2} \beta)^{-1} \|X_k S_k\|_\infty. \quad (39) \]

Then, it holds
\[ \|S_{k+1}^{-1}\|_\infty \leq C_0 (X_k S_k)^{-1}\|_\infty, \quad (44) \]
where \( C_0 = O(1) \).

Summarizing the above mentioned estimations (i.e., from (27), (38), (39) and (44)), we have
\[
\kappa(S_{k+1}) = \|S_{k+1}\|_\infty \|S_{k+1}^{-1}\|_\infty \\
\leq \|(X_k S_k) + \epsilon\|_\infty \|X_k S_k\|_\infty (X_k S_k)^{-1}\|_\infty \\
\leq (1 + C_7 + \epsilon) C_0 (1 + C'_1 + c_k C_5) \sqrt{\mu}(S_k) \\
\leq \mu_k \sqrt{\mu}(S_k) + O(1). \quad (45)
\]

Here, \( \mu_k := C_0 (C'_1 + c_k C_5)(1 + C_7 + \epsilon) = O(n) \).

Summing up the above mentioned discussions, we have the following theorem:

**Theorem 1** Assume that \( \kappa(S_k) \geq u^{-1} \). Further, let us assume that Assumptions 1, 3, 4 and 5 are satisfied. Then, \( \kappa(S_{k+1}) \leq \mu_k \sqrt{\mu}(S_k) + O(1) \) with \( \mu_k = O(n) \).

If \( \mu_k \sqrt{\mu} < 1 \) holds for \( k = 1, 2, \ldots, K \), then \( \kappa(S_k) \) decreases as \( O((n \sqrt{\mu})^k \mu_k) \) during \( k \leq K \) and finally \( \kappa(S_k) \) becomes \( O(1) \) as \( k \) becomes sufficiently large provided that \( k \leq K \).

**References**


Fast Verification for Sparse Linear Systems with Generalized Diagonally Dominant Matrices

Takeshi Ogita∗† and Shin’ichi Oishi†

∗ CREST, Japan Science and Technology Agency
† Faculty of Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
Email: {ogita, oishi}@waseda.jp

Abstract—This paper is concerned with the problem of verifying the accuracy of a numerical solution of a sparse linear system whose coefficient is a generalized diagonally dominant matrix (GDDM). A fast method is proposed for calculating an error bound of the numerical solution. The proposed method is based on the verified criterion for a GDDM and the rounding mode controlled verified computations. The method can be applied with any iterative solution methods such as Gauss-Seidel method, conjugate gradient method and so forth. Therefore, the sparsity of the coefficient matrix is not destroyed in the verification process. Numerical results are presented showing the validity of the proposed method.

1. Introduction

Let $A$ be a real $n \times n$ matrix and $b$ be a real $n$-vector. For a linear system

$$Ax = b,$$  

we can obtain a computed solution $\hat{x}$ of (1) using floating-point arithmetic. In general, however, we do not know how accurate the computed solution is because floating-point arithmetic causes rounding error.

Numerical methods of solving (1) can roughly be classified into two approaches. One is direct method (e.g. Gaussian elimination with partial pivoting) and the other is iterative method (e.g. Gauss-Seidel method, Krylov subspace method). To verify an accuracy of $\hat{x}$, several fast verification methods (e.g., [9, 11]) have been developed. Most of them are based on the direct method because their methods require an approximate inverse of $A$ or a lower bound of the smallest singular value of $A$.

The verification for sparse linear systems is still difficult except for the case where we know in advance that the coefficient matrix $A$ belongs to a certain special matrix class, e.g., M-matrix [7]. The difficulty is mainly due to the destruction of its sparsity in the verification process. Thus the verification for sparse systems becomes one of the open problems in the Grand Challenges and Scientific Standards in Interval Analysis [5] presented by Neumaier.

In this paper, we restrict ourselves to the case where $A$ is a generalized diagonally dominant matrix (GDDM). GDDDM has wide applications in engineering and scientific computations [1]. Here, we assume that we do not have any information on the property of $A$. That means it is necessary to prove whether $A$ is a GDDDM or not.

A main point of this paper is to develop a fast validation method in which computations are executed keeping the sparsity of matrix $A$. Thus, we do not calculate $A^{-1}$ nor an LU factorization of $A$. For the purpose, we use iteration algorithms in the verification process.

Numerical results are presented showing the validity of the proposed method.

2. Notation and Definitions

In this section, definitions of matrices and some known results used in this paper are in order. Throughout the paper, for a real matrix $A$, the notation $A \geq O$ ($A > O$) means that all elements of $A$ are nonnegative (positive), and for a real vector $v$, the notation $v \geq 0$ ($v > 0$) means that all elements of $v$ are nonnegative (positive).

Definition 1 Let $A = (a_{ij})$ be a real square matrix with $a_{ii} > 0$ and $a_{ij} \leq 0$ for $i \neq j$. Then $A$ is called an M-matrix if $A$ is nonsingular and $A^{-1} \geq O$.

Definition 2 Let $A = (a_{ij})$ be a real square matrix. Then a comparison matrix $M(A) = (\hat{a}_{ij})$ of $A$ is defined by

$$\hat{a}_{ij} = \begin{cases} |a_{ij}| & (i = j) \\ -|a_{ij}| & (i \neq j) \end{cases}.$$  

Definition 3 Let $A = (a_{ij})$ be a real $n \times n$ matrix. Let $N = \{1,2,\ldots,n\}$ and $N_i(A) = \{ i \mid |a_{ii}| > \sum_{j \neq i} |a_{ij}|, i \in N \}$. A matrix $A$ is called a strictly diagonally dominant matrix if $N_1 = N$.

Definition 4 Let $A = (a_{ij})$ be a real $n \times n$ matrix. A matrix $A$ is called a generalized diagonally dominant matrix (GDDDM) if there exists a positive diagonal matrix $D$ such that $AD$ is a strictly diagonally dominant matrix.

Lemma 1 $A = (a_{ij})$ is a GDDDM if and only if $M(A)$ is an M-matrix.

Lemma 2 $A = (a_{ij})$ is a GDDDM if and only if there exists a vector $u > 0$ such that $M(A)u > 0$.  

319
Lemma 3 If $A$ is a GDDM, then

$$|A^{-1}| \leq M(A)^{-1}.$$ 

From Lemma 3, it follows that

$$\|A^{-1}\|_\infty \leq \|M(A)^{-1}\|_\infty.$$  \hfill (2)

3. Verification Theory

We first cite a theorem for bounding $\|A^{-1}\|_\infty$ in case of $A$ being M-matrix, i.e. $A^{-1} \geq 0$.

Theorem 4 (Ogita-Oishi-Ushiro [7]) Let $A$ be a real $n \times n$ monotone matrix and $e$ be an $n$-vector where $e = (1, \ldots, 1)^T$. Let $\tilde{y}$ be a computed solution of a linear system $Ay = e$. Define a residual vector $s$ by $s := 1\tilde{y} - e$. If $||s||_\infty < 1$, then

$$\|A^{-1}\|_\infty \leq \frac{||\tilde{y}\|_\infty}{1 - ||s||_\infty}.$$ \hfill (3)

If $A$ is a GDDM, we can utilize the following corollary from (3) and Theorem 4:

Corollary 5 Let $A$ be a GDDM and $e$ be an $n$-vector where $e = (1, \ldots, 1)^T$. Let $\tilde{z}$ be a computed solution of a linear system $M(A)z = e$. Define a residual vector $t$ by $t := M(A)\tilde{z} - e$. If $||t||_\infty < 1$, then

$$\|A^{-1}\|_\infty \leq \frac{||\tilde{z}\|_\infty}{1 - ||t||_\infty}.$$ \hfill (4)

After bounding $\|A^{-1}\|_\infty$, we can calculate an error bound of an approximate solution $\hat{x}$ of $Ax = b$, for example,

$$\|\hat{x} - A^{-1}b\|_\infty \leq \|A^{-1}\|_\infty \|A\hat{x} - b\|_\infty \leq \frac{||\hat{x}\|_\infty}{1 - ||t||_\infty}.$$ \hfill (5)

There remains a problem how to prove whether $A$ is a GDDM or not. For the purpose, several methods (e.g. [2, 3, 4]) have been proposed. Using such methods, we can obtain a vector $u > 0$ which is expected to satisfy

$$M(A)u \geq 0.$$ \hfill (6)

In this paper, we use such a criterion for a GDDM and denote usage of the function as follows:

$$[\text{res}, u] = \text{isgddm}(A)$$

In view of the numerical verification, we have to confirm that $u$ satisfies (5). This is possible using rounding mode controlled computation [8, 10]. The following is an executable Matlab code for the purpose:

```matlab
function res = isgddm(A)
[res, u] = isgddm0(A); \% criterion for GDDM
B = compmat(A); \% comparison matrix
setround(-1) \% rounding downwards
v = B^u;
if all(v > 0)
    res = 1; \% verified A is a GDDM
else
    res = 0; \% not verified
end
```

4. Numerical Example

As a numerical example, we treat a test case in [2]:

$$A = \begin{bmatrix}
-1.0000 & -1.1463 & 0 & 0 & 0 \\
0.5000 & -1.0000 & 0 & -0.6000 & 0 \\
0 & -0.1000 & 1.0000 & 0 & 0.5000 \\
0 & 0.5000 & 0 & 1.0000 & -0.5000 \\
-0.2000 & 0.1000 & 0.3000 & 0 & -1.0000
\end{bmatrix}$$

This matrix is sensitive, i.e. if we set $A(1, 2) = -1.1464$, then $A$ is not a GDDM. In this test, we adopt the criterion of a GDDM in [2].

The following is a history of Matlab Command Window:

```matlab
>> [res, u] = isgddm0(A); res, u'
res = 1
ans =
1.0000 0.8723 0.2716 0.6206 0.3687
>> B = compmat(A)
B =
1.0000 -1.1463 0 0 0
-0.5000 1.0000 0 -0.6000 0
0 -0.1000 1.0000 0 -0.5000
0 -0.5000 0 1.0000 -0.5000
-0.2000 -0.1000 -0.3000 0 1.0000
>> setround(-1)
>> v = B^u; all(v > 0)
an =
1
>> setround(0)
>> e = ones(5, 1);
>> z = bicgstab(B, e);% bicgstab converged at iteration 4.5 to a solution with relative residual 8.5e-012
>> setround(-1)
>> t1 = B^z - e;
>> setround(1)
>> tu = B^z - e;
>> tu = max(abs(t1), abs(tu));
>> tnorm = norm(tu, inf)
tnorm = 1.4552e-011
>> invnorm = norm(z, inf) / (tnorm - 1)
invnorm = 5.7574e+004
>> norm(inv(A), inf)
an =
2.2788
```

From the above-mentioned result, an obtained upper bound $\text{invnorm}$ of $\|A^{-1}\|_\infty$ seems to be overestimated. However, it is not fatal because we can control the quality of an error bound of a computed solution [6].
References


Adaptive Verification Method for Dense Linear Systems

Katsuhisa Ozaki\textsuperscript{1}, Takeshi Ogita\textsuperscript{2,1} and Shin’ichi Oishi\textsuperscript{1}

\textsuperscript{1} Faculty of Science and Engineering, Waseda University, Tokyo 169-8555, Japan
\textsuperscript{2} CREST, Japan Science and Technology Agency (JST)

Email: k.ozaki@aoni.waseda.jp

Abstract—Several methods have been proposed to calculate an error bound of an approximate solution of a linear system by using floating-point arithmetic. It depends mainly on the condition number and the dimension of the coefficient matrix whether such verification methods succeed or not. In general, however, the condition number is not known in advance. If dimension or condition number is large enough, Oishi-Rump’s method, which is known as the fastest verification method for this purpose, may finish without success. There are more robust verification methods whose computational cost is larger than the Oishi-Rump’s one. However, it is not so efficient to apply such robust methods to well-conditioned problems. Therefore, we propose an adaptive verification method which automatically chooses an appropriate verification method with negligible cost, i.e. the proposed method does as much work as necessary to calculate error bounds of approximate solutions of linear systems.

1. Introduction

In this paper, we are concerned with a numerical verification method for a linear system

\[ Ax = b, \]

where \( A \) is a real \( n \times n \) matrix and \( b \) a real \( n \)-vector. A number of methods (e.g. \[3, 5, 6, 9\]) have been proposed for verifying an approximate solution of (1) with a dense coefficient matrix. Let \( R \) be an approximate inverse of \( A \) and \( \tilde{x} \) be an approximate solution of \( Ax = b \). Throughout the paper, we assume that an LU factorization with partial pivoting has been done such that

\[ PA \approx LU, \]

where \( L \) is a unit lower triangular matrix, \( U \) is an upper triangular matrix and \( P \) is a permutation matrix corresponding to the pivoting. Computation of an LU factorization requires \( \frac{2}{3}n^3 \) flops\footnote{addition, subtraction, multiplication or division are counted as one flop (floating-point operation)}. Utilizing the LU factors, computational cost for calculating \( R \) becomes \( 2n^3 \) flops.

Define \( \alpha \) and \( \beta \) by

\[ \alpha := \|RA - I\|_\infty \quad \text{and} \quad \beta := \|R(\tilde{x} - b)\|_\infty. \]

It is well-known that if \( \alpha < 1 \), then \( A \) is nonsingular and

\[ \|\tilde{x} - A^{-1}b\|_\infty \leq \frac{\beta}{1 - \alpha}. \] \hspace{1cm} (3)

Computational cost for calculating the error bound of \( \tilde{x} \) strongly depends on that for estimating \( \alpha \), which requires \( O(n^3) \) flops. Therefore, we focus our mind on the estimation of \( \alpha \). Until now, the following fast methods have been proposed using rounding mode controlled computation [6, 8]:

<table>
<thead>
<tr>
<th>Method</th>
<th>flops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oishi-Rump [6]</td>
<td>2/3 \cdot n^3</td>
</tr>
<tr>
<td>Ogita-Rump-Oishi [5]</td>
<td>4 \cdot n^3</td>
</tr>
<tr>
<td>Rump [8], Oishi-Rump [6]</td>
<td>6 \cdot n^3</td>
</tr>
</tbody>
</table>

There is a tradeoff between computational cost for the verification methods and treatable range of the problems. On the other hand, we see from (3) that any value \( \alpha \) less than 1/2, say, is sufficient for a reasonable error estimation. So not too much effort must be spent on estimating \( \alpha \).

In this paper, we consider to find adaptively which method is sufficient to prove \( \alpha < 1 \). A main point of the paper is to develop an adaptive verification algorithm which can automatically choose a suitable verification method corresponding to a given problem with a little additional cost.

2. Verification methods for linear systems

In this section, we briefly review fast verification methods for linear systems proposed in [5, 6].

Let \( \mathbb{F} \) be a set of floating-point numbers. Let \( \text{fl}(\cdot) \) be the result of a floating-point computations. Let \( u \) be the roundoff unit (especially, \( u = 2^{-53} \) in double precision defined in IEEE 754 standard). We define \( \gamma_n \) [2] by

\[ \gamma_n := \frac{nu}{1 - nu}. \]

First, we review the fastest method developed in [6]. Throughout the paper, we denote this method as Method-A. Let \( X_U \) and \( X_L \) be an approximate inverse of \( U \) and \( L \), respectively. Method-A computes \( X_U \) and \( X_L \), which requires \( \frac{1}{3}n^3 \) flops, respectively. By setting \( R := X_U X_L P \), Method-A estimates \( \alpha \) as

\[ \alpha = \|X_U X_L PA - I\|_\infty. \] \hspace{1cm} (4)
where $I$ denotes the $n \times n$ identity matrix. Define $e := (1, 1, \ldots, 1) \in \mathbb{R}^n$. Method-A evaluates an upper bound of $\alpha$ as

$$\alpha \leq \gamma_n(2||X_U| \cdot |X_L| \cdot |L| \cdot |U||\|_{\infty} + ||X_U| \cdot |U||\|_{\infty}).$$

(5)

Computation for the right-hand side of (5) can be done in $O(n^2)$ flops by using

$$||X| \cdot |Y||\|_{\infty} = ||X||(Y||e||)\|_{\infty},$$

where $X$ and $Y$ are real $n \times n$ matrices. The total computational cost for Method-A including computations of $X_U$ and $X_L$ becomes $2n^3$ flops. We denote the function based on Method-A as

$$[\text{err, flag}] = \text{Method-A}(\bar{x}, \bar{b}, L, U, P, X_L, X_U)$$

which returns an error bound $\text{err}$ of $\bar{x}$ and flag where $\text{flag} = 1$ means that the verification of $\bar{x}$ succeeded ($\alpha < 1$) and $\text{flag} = 0$ means that the verification failed ($\alpha \geq 1$).

Next, we review the method which is developed in [5]. Throughout the paper, we denote this method as Method-B. By Method-B, an upper bound for $\alpha$ is calculated using a priori error estimates as

$$\alpha \leq \text{fl}(||R_A - I||\|_{\infty}) + \gamma_{n+1}(||R| \cdot |A|\|\|_{\infty} + 1).$$

(6)

The method needs to compute an approximate inverse $R$ and matrix product $RA$. The total cost for Method-B including computations of $R$ and $RA$ becomes $4n^3$ flops. We denote the function based on Method-B as

$$[\text{err, flag}] = \text{Method-B}(\bar{x}, \bar{b}, R)$$

which returns similar outputs in Method-A.

We next review a method based on rounding mode controlled computation [6, 8]. Throughout the paper, we denote this method as Method-C. By Method-C, an upper bound of $\alpha$ is computed as follows:

$$\text{function} \ \alpha = \text{Alpha}(A, R)$$

setround(-1);

$Rd = RA - I$;

setround(1);

$Ru = RA - I$;

$$Ru = \max(|Ru|, |Rd|);$$

$$\alpha = \text{norm}(Ru, \text{inf});$$

Here, the instruction setround(-1) and setround(1) means to adopt the rounding-downwards mode and the rounding-upwards mode, respectively. Then, we assume that once the rounding mode is changed, it remains unchanged until the next instruction setround appears. This assumption is ensured on a wide class of computer systems following the IEEE 754 standard. Method-C requires $2n^3$ flops for calculating an approximate inverse $R$ and $4n^3$ flops for matrix product $RA$ in rounding-downwards and rounding-upwards. Therefore, the total cost for Method-C becomes $6n^3$ flops. We denote the function based on Method-C as follows:

$$[\text{err, flag}] = \text{Method-C}(\bar{x}, \bar{b}, R)$$

The above-mentioned verification methods can basically be applied to condition number of $A$ up to $\mathbf{u}^{-1}$. If we want to deal with more ill-conditioned problems, Rump’s method [7] can be used. Detailed algorithms and theory can be found in [7].

3. Adaptive Verification Method

In this section, we propose an adaptive verification method of obtaining an upper bound of $\alpha$ which automatically chooses a suitable verification method.

We first guess whether Method-A can be applied before calculating $X_U$. For the purpose, we utilize a 1-norm estimator developed by Higham [2] for estimating $||X_U||\|\|_{\infty}$. In fact, a lower bound of 1-norm and infinity-norm of an inverse of a triangular matrix can be estimated in $O(n^2)$ flops using forward or backward substitution. Throughout the paper, we denote the infinity-norm estimator for a triangular matrix $T$ as

$$c = \text{EstimateInv}(T).$$

In practice, the estimator usually gives a good estimation of $||T^{-1}||\|\|_{\infty}$. It is rare to underestimate the desired norm three times less than the actual norm. Using this estimator, we can estimate $||X_U||\|\|_{\infty}$ by $c = \text{EstimateInv}(U)$. Thus, we evaluate the right-hand side of (5) by

$$\gamma_n(2||X_U| \cdot |X_L| \cdot |L| \cdot |U||\|_{\infty} + ||X_U| \cdot |U||\|_{\infty})$$

$$\approx \gamma_n \cdot c||X_U| \cdot |L| \cdot |U||\|_{\infty} =: \alpha_A.$$  

The computation of $\alpha_A$ can be done in $O(n^2)$ flops. If $\alpha_A \geq 1$, then we proceed to Method-B without calculating $X_U$, i.e. we can save computational cost $\frac{1}{2}n^3$ flops in this case.

In Method-B, we can compute an approximate inverse $R$ of $A$ using $X_L$ and $U$ as

$$R = (U \setminus X_L)P,$$

where $X \setminus Y$ means to solve a matrix equation $XT = Y$ for $T$. We next guess whether Method-C can be applied to prove $\alpha < 1$. In (6), the term $||R| |A|\|\|_{\infty}$ can be calculated in $O(n^2)$ flops. If $||R| |A|\|\|_{\infty}$ then it is meaningless to compute the term $\text{fl}(||R_A - I||\|_{\infty})$. Moreover, numerical experiments show that

$$\text{fl}(||RA - I||\|_{\infty}) \approx \gamma_{n+1}||R| \cdot |A||\|_{\infty} \approx n\mathbf{u} \cdot \text{cond}\|A\|_\infty(A),$$

(7)

where $\text{cond}\|A\|_\infty(A) := ||A||_\infty||A^{-1}||_\infty$. Thus, we evaluate the right-hand side of (6) by

$$\text{fl}(||RA - I||\|_{\infty}) + \gamma_{n+1}||R| \cdot |A||\|_{\infty} + 1$$

$$\approx 2\gamma_n||R| \cdot |A||\|_{\infty} =: \alpha_B.$$
The computation of $\alpha_B$ can also be done in $O(n^2)$ flops. If $\alpha_B \geq 1$, then we proceed to Method-C without calculating $\text{fl}(RA)$, i.e. we can also save computational cost $2n^3$ flops in this case.

In Method-C, we can again use $R$. If Method-C cannot prove $\alpha < 1$, then it may be necessary to use higher precision computation because $A$ is ill-conditioned. In such case, we can use the Rump’s method for extremely ill-conditioned problems [7].

One may consider that it is natural to proceed as follows: we first compute $X_U$ and $X_L$ and estimate $\alpha$ using (5). After that if $\alpha < 1$ can not be proven, then we simply compute $R = X_U/X_LP$. Its computational cost is the same as that of calculating $R$ directly. From the view of optimization for computational speed, it is desirable to use routines supported by BLAS and LAPACK. Since the product of triangular matrices has not been supported in BLAS and LAPACK, it spends much more computing time than one may expect. This is the reason why we do not adopt this way.

Moreover, we consider the process of inverting $A$. As in the LAPACK’s routine DGETRI, an approximate inverse $R_1$ of $A$ using an LU factorization $PA \approx LU$ can be calculated by

$$X_U = I/L; \quad R_1 = (X_U/L)P;$$

where $X/Y$ means to solve a matrix equation $TY = X$ for $T$. There is an alternative way to calculate an approximate inverse $R_2$ of $A$:

$$X_L = I/L; \quad R_2 = (U/X_L)P;$$

One may also have a question how much the difference between $R_1$ and $R_2$ affects the quality of evaluating $\alpha$. To answer this question, we compare them in numerical experiment. Let $\alpha_1 := \text{fl}(R_1A - I)$ and $\alpha_2 := \text{fl}(R_2A - I)$. Let $\alpha_2$ and $\alpha_4$ be an upper bound of $\|R_1A - I\|_{\infty}$ and $\|R_2A - I\|_{\infty}$, respectively. Here, $\alpha_3$ and $\alpha_4$ are computed by the function Alpha in Section 2. We set $n = 1000$ and vary condition number of $A$ from 10 to $10^{13}$. We present a result of the numerical experiment in Table 1. From Table 1, it can be seen that there is a little difference between $\alpha_1$ and $\alpha_2$, which means that $R_1$ can produce a better result than $R_2$. However, there is not so much difference between $\alpha_2$ and $\alpha_4$. Therefore, in the sense of calculating an upper bound of $\|R_1A - I\|_{\infty}$, it seems to be not so wrong strategy to adopt the way of calculating $R_2$.

Combining the above-mentioned discussions, we finally present an adaptive algorithm in the following:

**Algorithm 1** Let $A \in \mathbb{R}^{n \times n}$ and $b \in \mathbb{R}^n$. The following algorithm returns an approximate solution $\tilde{x}$, its error bound $\text{err}$ and $\text{flag}$ which informs whether the verification succeeded or not, with choosing a suitable verification method.

**function** $[\tilde{x}, \text{err}, \text{flag}] = \text{OurMethod}(A, b)$

$[L, U, P] = \text{lup}(A);$ \quad \% LU factorization

<table>
<thead>
<tr>
<th>$\text{cond}(A)$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>$\alpha_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1.80e-12</td>
<td>3.91e-11</td>
<td>6.39e-10</td>
<td>6.49e-10</td>
</tr>
<tr>
<td>$10^5$</td>
<td>9.57e-10</td>
<td>1.20e-08</td>
<td>1.72e-07</td>
<td>1.92e-07</td>
</tr>
<tr>
<td>$10^6$</td>
<td>3.44e-06</td>
<td>5.74e-05</td>
<td>4.55e-04</td>
<td>5.73e-04</td>
</tr>
<tr>
<td>$10^{13}$</td>
<td>1.85e-02</td>
<td>3.74e-01</td>
<td>1.85e-02</td>
<td>4.93e-02</td>
</tr>
</tbody>
</table>

Of course, there is a possibility that $\alpha_{A,B} < 1$ while $\alpha \geq 1$, or $\alpha_{A,B} \geq 1$ while $\alpha < 1$. In such case, Algorithm 1 spends more computing time than a suitable verification method which is necessary or sufficient to prove $\alpha < 1$. In the next section, we will show that it is rare for such case to occur.

4. Numerical Examples

In this section, We report results of some numerical experiments to show the efficiency of the proposed method. The following four methods are implemented on a PC with Pentium IV 3.3GHz CPU and Matlab R14SP2:

A: Method-A ([6, Section 4])
B: Method-B ([5, Section 4])
C: Method-C ([6, Section 3])
D: Proposed method (Algorithm 1)

All computations are done in double precision defined in IEEE 754 standard on Matlab.

First, we set dimension $n = 1000$. Using Matlab’s function randsvd, we vary condition number of $A$ from 10 to $10^{14}$. We put $b = A \cdot e$. Table 2 displays elapsed time for each verification method. Here, the notation ‘−’ in the table means the verification failed. Note that elapsed time
Table 2: Comparison of elapsed time (sec) for dimension $n = 1000$.

<table>
<thead>
<tr>
<th>cond(A)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$10^1$</td>
<td>0.83</td>
<td>2.07</td>
<td>2.94</td>
<td>0.96</td>
</tr>
<tr>
<td>$10^2$</td>
<td>0.84</td>
<td>2.07</td>
<td>2.94</td>
<td>0.96</td>
</tr>
<tr>
<td>$10^3$</td>
<td>-</td>
<td>2.08</td>
<td>2.95</td>
<td>2.14</td>
</tr>
<tr>
<td>$10^{12}$</td>
<td>-</td>
<td>-</td>
<td>2.93</td>
<td>3.05</td>
</tr>
<tr>
<td>$10^{14}$</td>
<td>-</td>
<td>-</td>
<td>2.93</td>
<td>3.06</td>
</tr>
</tbody>
</table>

Table 3: Critical part of Table 2.

<table>
<thead>
<tr>
<th>cond(A)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>$8 \times 10^{12}$</td>
<td>-</td>
<td>-</td>
<td>2.93</td>
<td>3.06</td>
</tr>
<tr>
<td>$1 \times 10^{13}$</td>
<td>-</td>
<td>-</td>
<td>2.93</td>
<td>3.06</td>
</tr>
<tr>
<td>$2 \times 10^{13}$</td>
<td>-</td>
<td>-</td>
<td>2.93</td>
<td>-</td>
</tr>
<tr>
<td>$4 \times 10^{13}$</td>
<td>-</td>
<td>-</td>
<td>2.95</td>
<td>-</td>
</tr>
<tr>
<td>$6 \times 10^{13}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

for $A$ includes that for computing an approximate solution $\hat{x}$ using LU factors, and the elapsed time for $B$ and $C$ includes that for computing $\hat{x}$ by $x = Rb$. Each elapsed time includes that for the LU factorization.

Next, we set dimension $n = 3000$ and take $A$ and $b$ in the same way as in the previous example. The result of this example is displayed in Table 4.

From Tables 2 and 4, we can confirm that the proposed adaptive verification method $D$ can mostly choose a suitable verification method with a little additional cost. There are some cases where a suitable method is not chosen. For example, in Table 4, Method-A can deal with the problem of $\text{cond}(A) = 10^3$. However, the proposed method $D$ chooses Method-B.

To see more details in critical cases, we put Tables 3 and 5, which correspond to Tables 2 and 4, respectively. Table 3 shows that it occasionally occurs that Method-C succeeds while the proposed method $D$ does not. Moreover, Table 5 shows that there exists a case where it is difficult to determine which method is suitable, Method-A or Method-B. However, it can be seen that the range of such cases occurring is not so wide.

References


VSDP: A MATLAB software package for Verified Semidefinite Programming

Christian Jansson

Institute for Reliable Computing
Hamburg University of Technology
Schwarzenbergstr. 95, 21071 Hamburg, Germany
Email: jansson@tu-harburg.de

Abstract—VSDP is a MATLAB software package for solving rigorously semidefinite programming problems. Functions for computing verified forward error bounds of the true optimal value and verified certificates of feasibility and infeasibility are provided. All rounding errors due to floating point arithmetic are taken into account.

1. Introduction

Semidefinite Programming has emerged as a powerful tool in many different areas ranging from control engineering to structural design, combinatorial optimization and global optimization (see the Handbook of Semidefinite Programming [1]). One reason is that there exists a kind of calculus of conic quadratic and semidefinite representable sets and functions, which offers a systematic way to recognize and reformulate a convex program as a semidefinite program. This calculus is applied for example in CVX [2], an optimization modelling language which is designed to support the formulation and construction of optimization problems that the user intends from the outset to be convex. On the other hand non-convex problems are frequently solved by using convex relaxations, where consequently also SDP-solvers can be used.

Many algorithms for solving semidefinite programming problems require that appropriate rank conditions are fulfilled, and that strictly feasible solutions of the primal and the dual problem exist, i.e Slater’s constraint qualification holds. All these solvers do not provide a guaranteed accuracy or prove existence of optimal solutions. Nevertheless, appropriate warranties for computed results and rigorous forward error bounds can be useful in many situations, especially for ill-conditioned problems with dependencies in the input data, or ill-posed problems. It is well-known that for such problems (but not solely) rounding errors may affect the computation, and even many state-of-the-art solvers may produce erroneous approximations (cf. Neumaier and Shcherbina [3]).

Ill-conditioned and ill-posed problems are not rare in practice. In a paper of Ordóñez and Freund 2003 it is stated that 71% of the Ip-instances in the NETLIB Linear Programming Library are ill-posed, and recently Freund, Ordóñez and Toh 2006 [4] have shown that 32 out of 85 problems of the SDPLIB are ill-posed.

VSDP is a software package which provides warranties by computing verified forward error bounds. Verified, or sometimes also called rigorous, means that the computed results are claimed to be valid with mathematical certainty even in the presence of rounding errors due to floating point arithmetic. VSDP [5] is written completely in MATLAB under use of INTLAB [6]. It is based on a rigorous post-processing applied to the output of semidefinite programming solvers. It is of particular importance that each solver can be used, and the solver need not to produce any error bounds, neither in the forward nor in the backward error sense. This package implements techniques described in [7] and [8], and has several features: it computes verified lower and upper bounds of the optimal value for semidefinite programs, proves existence of feasible solutions, also for LMI’s, provides rigorous certificates of infeasibility, facilitates to solve approximately the problem by using different well-known semidefinite programming solvers, can handle several formats, and allows the use of interval data.

It is in the nature of verification methods that not every approximate solution can be verified, such as solvers normally cannot compute an approximate solution for each solvable problem. However, a good verification method should compute rigorous error bounds in almost all well-posed cases, whenever the used solver can compute a sufficiently close approximation. The numerical experiments of VSDP with the SDPLIB suite exhibit that at least for problems of middle size (up to thousands of constraints and millions of variables) rigorous lower (upper bounds) of the optimal value can be computed, provided the distance to dual infeasibility (primal infeasibility) is greater zero. But even if the distance to infeasibility is zero, i.e. the problem is ill-posed, VSDP allows rigorous results, if an a priori assumption about the existence of an optimal solution and its magnitude is known (c.f. [8]).

2. Quick Start

VSDP solves rigorously semidefinite programming problems in block diagonal form:

$$f_p := \min_{j=1}^n \sum_{j=1}^n \langle C_j, X_j \rangle \quad \text{s.t.} \sum_{j=1}^n \langle A_{ij}, X_j \rangle = b_i, \quad i = 1, \ldots, m$$

$$X_j \geq 0, \quad j = 1, \ldots, n.$$  

(1)
where \( b \in \mathbb{R}^m \), and \( C_j, A_{ij}, X_j \in S^s \), the linear space of real symmetric \( s_j \times s_j \) matrices. The usual inner product on the linear space of symmetric matrices is denoted by \( \langle \cdot, \cdot \rangle \), which is defined as the trace of the product of two matrices. \( X \succeq 0 \) means that \( X \) is positive semidefinite. Hence, \( \succeq \) denotes the Löwner partial order on this linear space. It is \( f^*_p := +\infty \) if the set of feasible solutions is empty, and \( f^*_p := -\infty \) if the problem is unbounded.

If \( s_j = 1 \) for \( j = 1, \ldots, n \) (i.e., \( C_j, A_{ij}, X_j \) are real numbers), then (1) defines the standard linear programming problem. The Lagrangian dual of (1) is

\[
f_d^* := \max b^T y \text{ s.t. } Z_j = C_j - \sum_{i=1}^m y A_{ij} \succeq 0 \text{ for } j = 1, \ldots, n,
\]

where \( y \in \mathbb{R}^m \). It is \( f_d^* := -\infty \), if the set of dual feasible solutions is empty, and \( f_d^* := +\infty \) in the unbounded case. The constraints \( \sum_{i=1}^m y A_{ij} \preceq C_j \) are called linear matrix inequalities (LMI’s).

Both problems are connected by weak duality

\[
f_d^* \leq f_p^*, \tag{3}
\]

but strong duality requires in contrast to linear programming additionally strict feasibility assumptions.

VSDP exploits the block-diagonal structure by an \( n \times 2 \) cell-array \( \text{blk} \), \( n \) cell-arrays \( C, X \), and an \( m \times n \) cell-array \( A \) as follows: The \( j \)-th block \( C[j] \) and the blocks \( A[i,j] \) for \( i = 1, \ldots, m \) are real symmetric matrices of common size \( s_j \) which is expressed by

\[
\text{blk}[j,1] = \textit{’s’}, \text{blk}[j,2] = s_j. \tag{1}
\]

The block-matrices \( C[j] \) and \( A[i,j] \) may be symmetric floating-point or interval matrices, and can be defined in dense or sparse format.

For the purpose of illustration, we start with the following semidefinite program of dimension \( m = 4, n = 1 \), and \( s_1 = 3 \), i.e. the matrices consist of only one block. The problem depends on a fixed parameter DELTA:

\[
>> \text{DELTA} = 1e^{-4};
\]

\[
>> C[1] = [ 0 \ 0 1/2 0; \ 1/2 \ 0 \ 0 \ \text{DELTA}; \ 0 \ 0 \ 0 \ 0 ];
\]

\[
>> A[1,1] = [ 0 \ -1/2 \ 0; \ -1/2 \ 0 \ 0; \ 0 \ 0 \ 0 ];
\]

\[
>> A[2,1] = [ 1 \ 0 \ 0; \ 0 \ 0 \ 0; \ 0 \ 0 \ 0 ];
\]

\[
>> A[3,1] = [ 0 \ 0 \ 1; \ 0 \ 0 \ 0; \ 0 \ 0 \ 0 ];
\]

\[1\]At the moment we have incorporated only symmetric matrices, which makes the first instruction redundant. But in future versions we want to distinguish also between other types of matrices. This structure is closely related to an older version of SDPT3.

It is easy to prove that this problem has a zero duality gap with the optimal value \(-0.5\) for every \( \text{DELTA} > 0 \). For \( \text{DELTA} = 0 \) the problem is ill-posed with nonzero duality gap, and for negative \( \text{DELTA} \) it is primal and dual infeasible. Especially, it follows that the optimal value is not continuous in \( \text{DELTA} = 0 \).

At the moment, the two semidefinite solvers SDPT3 and SDPA are adapted in a VSDP routine called MYSDPS. Therefore, VSDP can be used for computing approximations with different solvers. The user can integrate also other solvers very easily. By default, the function MYSDPS calls the semidefinite programming solver SDPT3:

\[
>> [\text{objt}, \text{Xt}, \text{yt}, \text{Zt}, \text{info}] = \text{mysdps(\text{blk}, A, C, b)};
\]

The output consists of approximations of (i) the primal and dual optimal value both stored in \( \text{objt} \), (ii) the primal and dual solutions \( \text{Xt, yt, Zt} \), and (iii) information about termination and performance stored in \( \text{info} \):

\[
>> \text{objt}, \text{termination} = \text{info(1)}, \text{objt} = -5.0003225608e-001 -5.0000000622e-001 \text{termination} = 0
\]

For \( \text{termination} = 0 \) we have normal termination without any warning. The first four decimal digits of the primal and dual optimal value are correct, but weak duality is not satisfied since the approximate primal optimal value is smaller than the dual one. In other words, the algorithm is not backward stable for this example. If we set the global variable \( \text{VSDP\_CHOICE\_SDP} = 2 \) in the file \( \text{SDP\_GLOBALPARAMETER} \), then the solver SDPA is chosen via \( \text{MYSDPS} \), and we obtain

\[
>> [\text{objt}, \text{Xt}, \text{yt}, \text{Zt}, \text{info}] = \text{mysdps(\text{blk}, A, C, b)};
\]

\[
>> \text{objt}, \text{termination} = \text{info(1)} \text{objt} = -8.4720539237e-0016.9954120952e-001 \text{termination} = 3
\]

No decimal digit of the optimal value is correct, but a warning is given, which indicates that the problem is primal or dual infeasible. To obtain more reliability we can use the function \( \text{VSDP\_LOW} \) which computes a verified lower bound of the primal optimal value by using a previously computed approximation. This function is based on the following theorem:
**Theorem 1** Assume that the maximal eigenvalues of a primal optimal solution \((X_j)\) are bounded by a nonnegative vector \(\lambda_{\min}(D_j)\) for \(j = 1, \ldots, n,\)

\[
D_j = C_j - \sum_{i=1}^{n} \hat{y}_i A_{ij}, \quad \text{and} \quad d_j := \lambda_{\min}(D_j) \quad \text{for} \quad j = 1, \ldots, n,
\]

where \(\lambda_{\min}\) denotes the smallest eigenvalue. Assume that \(D_j\) has at most \(l_j\) negative eigenvalues. Then the primal optimal value is bounded from below by

\[
f^*_p \geq b^T \hat{y} + \sum_{j=1}^{n} \frac{d_j^p}{2} \hat{x}_j =: f^*_p \quad \text{where} \quad d_j^p := \min(0, d_j).
\]

Moreover, if

\[
d_j^p \geq 0 \quad \text{for} \quad \hat{x}_j = +\infty,
\]

then the right hand side \(f^*_p\) is finite. If \(d_j^p \geq 0\) for \(j = 1, \ldots, n,\) then \(\hat{y}\) is dual feasible and \(f^*_p \geq \sum_{j=1}^{n} d_j^p\), and if moreover \(\hat{y}\) is optimal, then \(f^*_p = \sum_{j=1}^{n} d_j^p\).

There are no assumptions about the quality of \(\hat{y}\), but the last assertion implies that an approximation close to optimality should produce a rigorous lower bound with modest overestimation. The lower bound \(\sum_{j=1}^{n} d_j^p\) sums up the approximate dual value \(b^T \hat{y}\) and the violations of dual feasibility by taking into account the signs and multiplying these violations with appropriate primal weights.

**VSDPLOW** uses as starting point the already computed approximations \(X_t, y_t, Z_t\), and the call (all upper bounds \(\hat{x}_j\) are assumed to be infinite) has the form

\[
\gg \{F_L, Y, dL\} = \text{vsdp}low(blk, A, C, b, X_t, y_t, Z_t)
\]

The output \(F_L, Y, dL\) corresponds to the lower bound \(\sum_{j=1}^{n} d_j^p\), the certificate of dual feasibility \(Y\), and the vector of eigenvalue bounds \(d\) where \(\hat{y} = Y\), respectively. In the case where no certificate of feasibility could be computed we set \(Y = \text{NaN}\).

With the SDPA approximations the rigorous lower bound is infinite, and dual feasibility is not verified, i.e. \(Y = \text{NaN}\). But working with the SDPT3 approximations yields

\[
\gg \{F_L, Y, dL\} = \text{vsdp}low(blk, A, C, b, X_t, y_t, Z_t, xu)
\]

With SDPT3 a finite rigorous lower bound close to the optimal value together with a certificate of dual feasibility \(Y\) is computed. Therefore, strong duality is verified. If finite upper bounds \(xu = (\hat{x}_j)\) are known then the call of **VSDPLOW** is

\[
\gg \{F_L, Y, dL\} = \text{vsdp}low(blk, A, C, b, X_t, y_t, Z_t, xu)
\]

Similarly, with function **VSDPUP** we can compute a verified upper bound \(FU\) of the optimal value using the previously computed approximations. This function is based on the following theorem:

**Theorem 2** Assume that the absolute value of a dual optimal solution is bounded by a vector \(\tilde{y} > 0\), which may also have infinite components. Let \(X_j \in S^n\) for \(j = 1, \ldots, n,\) and assume that each \(X_j\) has at most \(k_j\) negative eigenvalues. Let for \(i = 1, \ldots, m\) and \(j = 1, \ldots, n,\)

\[
r_i \geq |b_i - \sum_{j=1}^{n} (A_{ij}, \hat{X}_j)|
\]

\[
\gamma_j \leq \lambda_{\min}(X_j), \quad \text{and}
\]

\[
\varrho_j \geq \sup\{\lambda_{\max}(C_j - \sum_{i=1}^{m} y_i A_{ij}) : -\tilde{y} \leq y \leq \tilde{y}, \quad C_j - \sum_{i=1}^{m} y_i A_{ij} \succeq 0\}
\]

Then the dual optimal value satisfies

\[
f^*_d \leq \sum_{j=1}^{n} (C_j, \hat{X}_j) - \sum_{j=1}^{n} k_j \varrho_j + \sum_{i=1}^{m} r_i \tilde{y}_i =: f^*_d,
\]

where \(\gamma_j := \min(0, \gamma_j)\). Moreover, if

\[
r_i = 0 \text{ for } \tilde{y}_i = +\infty \quad \text{and} \quad \gamma_j = +\infty,
\]

then the right hand side \(f^*_d\) is finite. If \(\gamma_j \geq 0\) and \(r_i = 0\) for all \(i, j, \) then \((X_j)\) is primal feasible and \(f^*_p \leq f^*_d\). If moreover \((X_j)\) is optimal, then \(f^*_p = f^*_d\).

The bound \(f^*_d\) sums up the approximate primal objective value \(\sum_{j=1}^{n} (C_j, \hat{X}_j)\) and the violations of primal feasibility \((r_i\) and \(\gamma_j)\) by taking into account the signs and multiplying these violations with appropriate weights \(r_i\) and \(\gamma_j\). The call of **VSDPUP** has the form

\[
\gg \{F_U, X, 1b\} = \text{vsdp}up(blk, A, C, b, X_t, y_t, Z_t);
\]

\[
F_U = -4.9996776932e-001
\]

The output \(F_U, X\) and \(1b\) corresponds to the upper bound \(f^*_d\), the interval block-diagonal matrix (containing the rigorous certificate of primal feasibility), and the vector of eigenvalue bounds \(\gamma_j\).

Summarizing, by using the SDPT3 approximations we have verified the inequality

\[-5.0000000662e-001 \leq f^*_d = -4.9996776932e-001,\]

Certificates of strictly primal and strictly dual feasible solutions are computed. The Strong Duality Theorem implies that the primal and the dual problem have a nonempty compact set of optimal solutions. The upper and lower bounds of the optimal value show a modest overestimation, mainly due to the accuracy of SDPT3.
Further numerical results for different values Delta are summarized in Tables 1 and 2. The approximate primal and dual optimal value computed by SDPT3 are denoted by \( \hat{f}_p \) and \( \hat{f}_d \), respectively. The value \( \hat{\rho} \) is the maximum of the relative gap and the measures for primal and dual infeasibility. In all cases the default values of SDPT3 are used, and normal termination without warning has occurred. SDPT3 is not backward stable, since in two cases \( \hat{f}_p < \hat{f}_d \) violating the weak duality. For the smallest value of Delta no decimal digit of \( \hat{f}_p \) or \( \hat{f}_d \) is correct. In all cases no warning was given. The approximate residual \( \hat{\rho} \) leads to the suspicion that at least five decimal digits are correct. The new rigorous bounds (which use the computed approximations of SDPT3) reflects much more the reliability of SDPT3, and the measures for primal and dual infeasibility. For the smallest value of Delta no decimal digit of \( \hat{f}_p \) or \( \hat{f}_d \) is correct. In all cases no warning was given. The approximate residual \( \hat{\rho} \) leads to the suspicion that at least five decimal digits are correct. The new rigorous bounds (which use the computed approximations of SDPT3) reflects much more the reliability of SDPT3, and the number of correct decimal digits for the computed result. The bounds FU and FL fulfill weak duality, and the true optimal value \( -1/2 \) is inside the bounds, which is not the case for the approximations \( \hat{f}_p \) and \( \hat{f}_d \) corresponding to the values Delta = \( 10^{-4} \) and Delta = \( 10^{-5} \).

### 3. Rigorous Error Bounds for the SDPLIB

The SDPLIB is a collection of semidefinite programming problems with different areas of applications. Freund, Ordoñez and Toh [4] have solved 85 problems of the SDPLIB with SDPT3. They have shown that 32 are ill-posed. VSDP could compute (by using SDPT3 as approximate solver) for all 85 problems a rigorous lower bound of the optimal value and verify the existence of strictly dual feasible solutions. This implies a zero duality gap for all these problems. A finite rigorous upper bound could be computed for all well-posed problems with one exception; this is hinf8 being ill-conditioned. For all 32 ill-posed problems VSDP has computed \( \hat{f}_d = +\infty \), which reflects exactly that the distance to the next primal infeasible problem is zero as well as the infinite condition number.

Detailed numerical results can be found in [5]. For the 85 test problems, SDPT3 (with default values) gave 32 warnings, but 13 warnings were given for well-posed problems. No warning was given for 13 ill-posed problems. In other words, there is no correlation between warnings and the difficulty of the problem. What is the sense of warnings? I have no satisfactory answer. But rigorous bounds provide safety and are important, especially in the case where algorithms subsequently call other algorithms, as is done for example in branch-and-bound methods.

Some major characteristics of our numerical results for the SDPLIB are as follows: The median of the time ratio for computing the rigorous lower (upper) bound and the approximation is 0.045, (2.4), respectively. The median of the guaranteed accuracy for the problems with finite condition number is \( 4.9 \cdot 10^{-7} \). We have used here the median because there are some outliers. One of the largest problems which could be solved by VSDP is thetaG51 where the number of constraints is \( m = 6910 \), and the dimension of the primal symmetric matrix \( X \) is \( s = 1001 \) (implying 501501 variables). For this problem SDPT3 gave the message out of memory, and we used SDPA as approximate solver. The rigorous lower and upper bounds computed by VSDP are \( \text{FL} = -3.4900 \cdot 10^{2}, \text{FU} = -3.4406 \cdot 10^{2} \), respectively. This is an outlier because the guaranteed relative accuracy is only 0.014, which may be sufficient in several applications, but is insufficient from a numerical point of view. However, existence of optimal solutions and strong duality is proved. The times in seconds for computing the approximations, the lower and the upper bound of the optimal value are \( t = 3687.95, tFL = 45.17, \text{and } tFU = 6592.52 \), respectively.

### References


Fractal with price scale of high-frequency data in futures market

Yoshiaki Kumagai†

†Faculty of Education and Integrated Arts and Sciences, Waseda University
1-6-1, NishiWaseda, Shinjuku-ku, Tokyo 169-8050, Japan
Email: kumagai-y@waseda.jp

Abstract—The fractal structure of high-frequency financial data has been analyzed by the folding dimension, which has no relation with time scale. The folding dimension and Kagi chart which is one of Japanese charting techniques are based on the same concept of scaling along the price axis. The data set used is the gasoline futures in the Tokyo Commodity Exchange (TOCOM) in 2005. The daily fractal structure of the futures prices is measured by the folding dimension of tick-by-tick data. The fractal dimension in the futures market showed time-variability. However, in most of the days the futures prices are quasi-persistent in contract to foreign exchange rates which are strong quasi-anti-persistent.

1. Introduction

Transactions in financial markets occur at varying time intervals. Most studies, however, have converted such data into that of equal intervals. High-frequency data are often re-sampled at intervals of 5 or 10 minutes. Moreover, “time”, which evolves the price, cannot be determined uniquely. There are at least four candidates for the proper time scale to use for analyzing market data: physical time, trading time, the number of trade, and cumulated trading volume [1]. We propose to use another simple method to describe the fractal structure of data, which are irregularly spaced by nature.

First we expand the definition of extreme values, to be determined with a scale that corresponds the price-axis resolution. Using these extreme values determined with a scale, we can measure a kind of the fractal dimension: the fold dimension. When using the ordinary methods of fractal analysis, the least division of the time scale of these data is one minute. Therefore, the rest of information from the representative of one-minute interval is not used. On the other hand, by using the fold dimension $D_f$, we can analyze the data without information losses but time intervals [7]. This dimension captures the scaling behavior of the time series without loss of information along with the price-axis.

Next we show that this method is related to technical analysis. The folding dimension is based on the same idea as a Japanese charting technique: Kagi chart.

In this paper, we use a high-frequency data set on gasoline futures prices in the Tokyo Commodity Exchange (TOCOM).

2. Scaling of fluctuations

The local maximum determined by a scale $C$ was defined as follows [5]. Suppose $g$ is a real function defined on $\mathbb{R}^1$. We say that $g$ has a local maximum determined by a scale $C$ at point $t$, if there exist $d_1, d_2 > 0$ such that $g(t+d_2) > g(t)$ and $g$ attains its maximum at $t$ in $[t-d_1, t+d_2]$. The local maximum determined by a scale $0$ means the ordinary local maximum. The width $C$ is the accuracy of measurement of extreme values. The local minimum determined by a scale $C$ is defined likewise.

Using these extreme values, we define two functions. In this section, suppose that the time series $g(t)$ is defined in a given interval $[0, S]$. Let $C$ denote a vertical spread. In this interval $[0, S]$, we can detect extreme values determined by a scale $C$. Suppose $g(t)$ attains its extreme values at $t_1^C < t_2^C < t_3^C < \cdots$, ordered increasingly. With $g(t)$ fixed, the number of these maxima and minima in $[0, S]$ is a function of $C$. Let $m(C)$ denote the number of these extreme values. Next, we introduce a kind of total variation of the time series.

Let $R(C)$ denote the total sum of absolute variation between neighboring extreme values:

$$R(C) \equiv \sum_{i=1}^{m(C)-1} |g(t_{i+1}^C) - g(t_i^C)|. \quad (1)$$

$R(0)$ is the total sum of absolute variations: $R(0) = \lim_{n \to \infty} \sum_{k=1}^{n} |g((k-1)S/n) - g(kS/n)|$. Note that $m(C)$ and $R(C)$ are invariants for the choice of time scale. The information used is the values and their order in time, not the time. This property is significant to investigate non-equidistant data like high-frequency data of price series. Further properties of $m(C)$ and $R(C)$ are in [5].

Using the function $R(C)$, a kind of fractal dimension was defined [6]. If the relation:

$$R(C)/C \sim C^{-D_f}.$$  \quad (2)
Figure 1: Example of extreme values determined by a scale \( C \). The dashed line denotes a function \( g(s) \). The function \( g \) attains its local maximum determined by a scale \( C \) at \( t_1 \), but \( g \) does not attain its maximum at \( t_{01} \). Using a slit: a pair of parallels whose distance is \( C \) (denoted by short dashed lines), we can detect these extreme values.

On the other hand, \( g \) attains its local minimum with \( C \) at \( t_2 \), but \( g \) does not attain its minimum at \( t_{02} \).

holds, we say that the exponent \( D_f \) is the fold dimension of the time series. When the time series chart is self-affine, the fold dimension \( D_f \) is correspond to the latent fractal dimension \( D_l \). The relation between the local box dimension \( D_b \) and the latent fractal dimension \( D_l \) is \( D_b = 2 - 1/D_l \) \cite{10}. The Hurst exponent \( H \) is expressed as \( H = 2 - D_b \) \cite{4}. Therefore we call the price series is "quasi-persistent" when \( 2-1/D_f < 1.5 \), and "quasi-anti-persistent" when \( 2-1/D_f > 1.5 \).

3. Kagi charts and scaling

The charts in the technical analysis of the trend type can be categorized by the \( x \)-axis: time series charts and non-time series charts. Rather than having price on the \( y \)-axis and time on the \( x \)-axis, non-time series charts disregard the passage of time and display price changes on both axes. This category includes Kagi, Renko, Three Lines Break and Point & Figure(P&F) charts.

Kagi charts display a series of connecting vertical lines where the direction and the thickness (or color) of the lines are dependent on the price action. The charts ignore the passage of time. If prices continue to move in the same direction, the vertical line is extended. However, if prices reverse by a "reversal" amount \( C \), a new kagi line is then drawn in a new column. When prices penetrate a previous high or low, the thickness of the kagi line changes. There are many trading techniques using Kagi chart, but the most basic trading technique is to buy when the kagi line changes from thin to thick and to sell when the kagi line changes from thick to thin.

We now show the algorithm to draw Kagi chart for the price series \( p(s), (s = 0, 1, 2 \cdots) \) using the reversal amount \( C \). Let \( z^- \) denote the candidate for the upward reversal, and \( z^+ \) the candidate for the downward reversal. The function \( q \) is an indicator which represents whether the line extend upwards or downwards at time \( s \). The calculations at time \( s \) are as follows.

If \( p(s) > p(s-1) \) then
  If \( p(s) > z^+ \) then
    \( z^+ \leftarrow p(s) \)
    The line is extended in the same direction (upwards).
  If \( p(s) < z^- \) then
    \( z^- \leftarrow p(s) \)
    The line is extended in the same direction (downwards).

The reversal amount \( C \) is fixed of the size historically used for that particular price range or the size by a constant percentage of the price. It is important to find an appropriate reversal amount for the price series. An example of Kagi chart is shown in Fig. 2. For more details on these charts, see \cite{11}.

Kagi charts and extreme values determined by a scale are based on the same concept of scaling along the \( y \)-axis. Local maxima determined by a scale \( C \) correspond to the upper ends of the vertical lines in Kagi charts drawn by the reversal amount \( C \). Likewise, local minima determined by a scale \( C \) correspond to the upper ends of the vertical lines in Kagi charts drawn by the reversal amount \( C \). Hence the number of the vertical lines in \([0, S]\) is equal to the number of extreme-values \( m(C) \). Furthermore, the sum of the length of vertical lines is equal to the total variation \( R(C) \). Making \( C \) larger filters out more data. The chartists control the reversal amount \( C \) to find a balance that shows enough detail without showing too much. On the other hand, we control \( C \) to detect the fractal structure of the time series.
Figure 2: Kagi chart of NEC in Tokyo Stock Exchange from Jun. 1 to Nov. 30, 1999, drawn by the reversal amount 100 yen. The dotted line represents a bar chart of daily closing prices for the same period (referring the upper axis). The figure does not show the thickness, since it is not related to the scaling of the time series.

4. Data

The data set used is gasoline futures\(^1\) in the Tokyo Commodity Exchange (TOCOM) from Jan. 1, 2005 to Dec. 31, 2005. In TOCOM, six contract months of gasoline are traded at any time. The last trading day is the 25th of the month preceding the delivery month.\(^2\) The number of ticks and the trading volume is large in the distant futures and decrease when a new contract starts. Fig. 3 shows two consecutive contract months of Sep. and Oct. in 2005. The number of ticks and the trading volume of the distant contracts of gasoline futures prices in TOCOM are shown in Fig. 4.

5. Results

The efficient slopes of log \(R(C)/C\) with log \(C\) dependent on the scaling regions. Here the scaling region is 10-150 Yen.

The evolution of daily dimensions \((2−1/D_f)\) and the gasoline futures prices are shown in Fig. 5. The day when the number of ticks is less than 3000 is excluded.

In most days of the year, the futures prices are quasi-persistent. This result is similar to the ordinary fractal analysis on high frequency data in finance, in which the price series are found to be persistent: \(H > 0.5\) [2] [3]. However this is contrast to the result for foreign exchange market [8]. The quotes (bid and ask) for US Dollar-Japanese Yen exchange rates submitted by the

\(^1\) The standard is regular gasoline of JIS K2202 Grade 2 with maximum permissible sulfur content of 10 ppm. The contract unit is 50 kl. The minimum fluctuation is 10 Yen per 1 kiloliter.

\(^2\) If the day is a holiday, last trading day is advanced.
brokers are quasi-anti-persistent from Jan. 1, 1995 to Nov. 30, 2000. Also using other ordinary fractal analysis, anti-persistency: $H < 0.5$ is reported on time scale up to several tens of minutes in foreign exchange market [12]. With longer time scale, exchange rates are reported to be random series: $H = 0.5$ [9]. In the foreign exchange market, standard deviation of the price changes (tick time) and the fractal dimension $2 - 1/D_f$ have negative correlation. This relation also hold in the gasoline futures market. Thus when the market becomes volatile in this standard deviation, the quasi-anti-persistence weakens in exchange rates, whereas quasi-persistence strengthens in futures.

6. Conclusion

In this paper, we use a time-scale independent method to describe the fractal structure of data, which are irregularly spaced by nature. This method is the revision of the traders’ perception of price series from econophysics. The daily fractal structure of futures prices is measured by the folding dimension of tick-by-tick data. The fractal dimension in the futures market showed time-variability. However, in most of the days prices are quasi-persistent in contract to foreign exchange rates.

Acknowledgements

We acknowledge the continuing guidance and encouragement of Gyoichi Iwata, Toshio Inaba and Yuki-tami Tsuji.

References


Monetary policy subject to measurement errors of private sector adaptive expectations

Stefano Casaccia, Silvano Cincotti*, Marco Raberto, Andrea Teglio

DIBE-CINEF, University of Genova
Via Opera Pia 11a, Genova 16145, Italy
*Corresponding author: cincotti@dibe.unige.it

Abstract—We study optimal monetary policy by performing different computational experiments within the framework of two-equation stochastic model of the economy with a forward-looking private sector. The private sector forms adaptive expectations on inflation and output gap. The central bank sets the interest rate in order to minimize a loss function depending both on inflation and output gap. The monetary authority takes into account non-rational expectations by the private sector in order to set the optimal interest rate. However, measurements of private sector expectations are affected by noise and measurement errors critically affect the ability of the policy maker to control of the economy. The main finding of this study is that the performance of the monetary policy depends critically more on measurement errors of inflation expectations than of output gap expectations.

1. The macroeconomic framework

1.1. The two-equation model

The formulation and performance of monetary policy has been analyzed extensively in the recent literature [1, 2]. This study addresses this issues within a baseline framework of a forward looking, two-equation model for inflation and output gap which is standard in the literature; see [1] and reference therein. The model is a log-linearized approximation of a dynamic general equilibrium nonlinear sticky price model. The log-linear approximation is taken around the non stochastic steady state with zero inflation and steady state output growth, see Walsh [3, Cap 5] for details on the derivation. The nonlinear framework is based on the optimizing behavior of a representative consumer and of a continuum of firms producing goods under monopolistic competition and subject to staggered price changes.

The model framework consists of two equations:

\[ x_t = -\phi(i_t - E^*_t\pi_{t+1}) + E^*_t x_{t+1} + g_t, \]  \( E^*_t \) and \( \pi_{t+1} \) are the density of the observable real interest rate and the expected inflation for the next period, respectively. The parameter \( \phi \) represents the sensitivity of the output gap to changes in the short-term interest rate. The parameter \( \lambda \) represents the sensitivity of inflation to changes in the real interest rate. The parameters \( \beta \) and \( \lambda \) depend on the structural parameters of the nonlinear optimizing framework, and is the inverse of the intertemporal elasticity of substitution of consumption today with consumption tomorrow by the representative household.

Equation 1 represents the demand side of the economy and represents a dynamic, forward-looking IS curve which can be derived from the log-linear approximation of the consumption Euler equation arising from the optimal household saving decision, after imposing the equilibrium condition in the goods market. According to Eq. 1, output depends on expected future output \( E^*_t x_{t+1} \) as well as the real interest rate \( E^*_t \pi_{t+1} \). The negative effect of the real rate on current output reflects the intertemporal substitution of consumption. In this respect, the parameter \( \phi \) has a direct interpretation based on the structural parameters of the nonlinear optimizing framework, and is the inverse of the intertemporal elasticity of substitution of consumption today with consumption tomorrow by the representative household.

1.2. Economic interpretation

Equation 2, often referred as the new Keynesian Phillips curve, corresponds to the supply side of the economy and is derived from the optimizing behavior of firms, conditional on the assumed economic environment (monopolistic competition, constant elasticity demand curve, and staggered nominal price setting). In particular, Eq. 2 is a log-linear approximation around the steady state of the aggregation of firms pricing decision. Parameters \( \beta \) and \( \lambda \) depend on the structural parameters of the nonlinear optimizing frame-
work. $\beta$ is the household discount factor of future utility and $\lambda$ depends both on $\beta$ and on the probability of price adjustment and can be interpreted as the output elasticity of real marginal costs. Indeed, in this framework, marginal costs are proportionate to the output gap and, in this respect, shocks $u_t$ are random deviations from this proportionality conditions caused, for example, by movements in nominal wages that push real wages away from their equilibrium values due to frictions in the wage contracting process.

2. Discretionary monetary policy

According to the literature on flexible inflation targeting, see [4], we assume that the central bank sets the nominal interest rate $i_t$ in order to minimize a quadratic loss function $L$ defined in terms of output gap and inflation, i.e.,

$$ L = \sum_{s=0}^{\infty} \beta^s [\alpha x_t^2 + \pi_t^2],$$

where $\beta$ is the discount factor assumed to be the same for the private sector and $\alpha$ is the relative weight placed by the policy maker on output deviations. The case $\alpha = 0$ refers to a strict inflation targeting monetary policy. The literature on optimal policy distinguishes between optimal discretionary policy, in which the policy maker is unable to commit to policies for future periods, and optimal policy in which such commitment is possible. This paper focuses on discretionary monetary policy. In such case, policy is re-optimized each period and reduces to a sequence of static problems in which the Central Bank aims to minimize $\alpha x_t^2 + \pi_t^2$ subject to Eq. 2. This leads to a constraint condition in output gap and inflation rate, i.e.,

$$ \lambda \pi_t + \alpha (x_t - x_{t-1}) = 0. \quad (4)$$

2.1. Rational expectations equilibrium solution

Assuming rational expectations (RE) by private agents, it can be shown, see [5], that the dynamic system in $x_t$ and $\pi_t$ defined by Eq. 2 and Eq. 4 has a unique nonexplosive rational expectation solution, i.e.,

$$ x_t = c_x u_t \quad \text{with} \quad c_x = -\frac{\lambda}{\alpha(1 - \beta \rho) + \lambda^2} \quad (5)$$

$$ \pi_t = c_\pi u_t \quad \text{with} \quad c_\pi = -\frac{\alpha}{\lambda} c_x. \quad (6)$$

We will refer to Eq. 5 and Eq. 6 as the rational expectation equilibrium (REE) solution. Under RE, the optimal interest rate policy is straightforward. According to Eq. 5 and Eq. 6, RE $E_t x_{t+1}$ and $E_t \pi_{t+1}$ are given by:

$$ E_t x_{t+1} = c_x \rho u_t, \quad (7)$$

$$ E_t \pi_{t+1} = c_\pi \rho u_t. \quad (8)$$

Inserting Eq. 7, 8 and Eq. 5 into Eq. 1, and solving for $i_t$, the optimal interest rate rule is then given by:

$$ i_t = \phi^{-1} g_t + (\phi^{-1} c_x (\rho - 1) + c_\pi \rho) u_t. \quad (9)$$

We will refer to Eq. 9 as the fundamentals-based interest rate policy as the interest rate rule is formulated in terms of observed fundamentals shocks $g_t$ and $u_t$ to the economy.

2.2. Learning adaptive expectations

The policy rule of Eq. 9 is based on the assumption that the economy is in a stationary REE, i.e., presupposes REE
both on the part of private agents and the policy maker. Recently, Evans and Honkapohja [5] showed that the economy modelled by Eqs. 1 and 2 is unstable if the policy maker follows the fundamentals-based interest rate policy and private agents follow adaptive learning rule to form expectation of future inflation and output gap. Small deviation of adaptive expectations from the rational ones become magnified by the policy, which assumes RE, and the cumulative process drives the economy away from the rational expectations equilibrium. Evans and Honkapohja [5] derived a new optimal interest rate policy, called the expectations-based interest rate rule, i.e., Eq. 10. Finally, according to the fundamentals-based policy, i.e, Eq. 9, or the expectations-based interest rate policy for any value of the structural parameters of the model. Conversely, under the fundamentals-based interest rate, the system does not converge to the REE for any value of the structural parameters.

An important issue regards how possible measurement errors on the observations of current expectations of future variables may affect the performance of monetary policy. Survey data on private forecasts of future inflation and various measures of future output do exist but there are concerns about the accuracy of these data. If measurements errors are large, a substantial deterioration in the performance should be expected. The simulation results discussed in the next section are mainly focused on this topic.

3. Computational experiments

Different computational experiments have been performed simulating the system described by Eqs. 1 and 2 with the interest rate $i_t$ set according to Eq. 9 or Eq. 10. The values of structural parameters have been set according to the calibration by Clarida et al. [6] on real data, i.e., $\beta = 0.99$, $\phi = 4$, and $\lambda = 0.075$. We also set the parameters of the random shocks as $\mu = 0.5$, $\rho = 0.5$, and $\sigma_u = 1$, unless otherwise stated.

Figures 1 and 2 present the dynamics of the output gap $x_t$ and $\pi_t$ after an inflation shock, with the value of weight $\alpha$ set to 0.25. The values of $x_t$ and $\pi_t$ are initially set to their steady-state long-run values. The steady state dynamics has been simulated considering two initial ran-
dom shocks to the economy, i.e., \( \tilde{g}_t = N(0, \sigma_{\tilde{g}}^2) \delta(t) \) and \( u_t = N(0, \sigma_{u}^2) \delta(t) \). It is worth noting that the steady state values do not coincide here with the REE solution, as shown in the Figures. The bias depends on the persistence effect of initial random shocks due to non-zero values of \( \mu \) and \( \rho \). At \( t = 50 \), a positive inflation shock is applied to the economy, then the dynamics of \( x_t \) and \( \pi_t \) is displayed considering the two interest rate policies discussed in the previous section. Results clearly outline the different capabilities of two policies to control the economy and confirm previous theoretical finding by Evans and Honkapohja. Under the fundamentals-based policy, the economy diverges from its previous steady-state dynamics, whereas under the expectations based control rule, the values of \( x_t \) and \( \pi_t \) return back to their steady state values.

In order to evaluate the impact of measurements errors by the policy maker, we have performed different simulations under the expectations-based policy, considering multiplicative white noise errors on expectations \( E^*_t x_{t+1} \) and \( E^*_t \pi_{t+1} \). Errors are normally distributed with mean 1 and standard deviation \( \sigma_{E^*_x} \) and \( \sigma_{E^*_\pi} \), respectively. Different values of \( \alpha \) have been considered in the range between 0.05 and 1. Results are showed in Fig. 3 and Fig. 4 for the six cases considered. The REE solution, absence of errors on expectations observations (\( \sigma_{E^*_x}=\sigma_{E^*_\pi}=0 \)), and errors on the measurement of inflation and output gap expectations, are considered separately. On the side of output gap expectations errors, we have considered two cases with \( \sigma_{E^*_x} = 0.4 \) and \( \sigma_{E^*_x} = 0.6 \), respectively. As regarding errors on inflation expectations, we have two cases with \( \sigma_{E^*_\pi} = 0.1 \) and \( \sigma_{E^*_\pi} = 0.15 \), respectively.

Figure 3 shows the efficient frontier occurring from the tradeoff between the average of the square of the output gap, i.e. \( E(x^2) \), and the average of the square of the inflation rate, i.e. \( E(\pi^2) \). Figure 4 displays the estimated values of the loss function, see Eq. 3, for the six cases considered. The two figures point out a major result, i.e., the system under RLS learning and with the expectations-based interest rate rule is controlled better than in the REE solution case with the fundamentals-based policy. Measurements errors on the output gap expectations deteriorate the performance. However, with the values of \( \sigma_{E^*_x} \) considered, the loss function results to be lower than in the REE case. Conversely, even small observation errors on the side of inflation expectations, worse significatively the capability of the monetary authority to control the economy, with an overall performance lower than the REE case.

4. Concluding remarks

This study presented computational experiments on optimal monetary policy evaluation within the framework of a forward looking model of the economy with adaptive expectations by the private sector. The main result of the study is that monetary authority should take into account agents’ expectations in setting its policy and that even small observation errors on inflation expectations may deteriorate significatively the capability to control the economy.

Acknowledgments

This work has been partially supported by the University of Genoa, and by the Italian Ministry of Education, University and Research (MIUR) under grants FIRB 2001 and COFIN 2004, and by the European Union under IST-FET STREP Project EURACE.

References

Nonconvex Variational Problem with Recursive Integral Functionals in Sobolev Spaces: Existence and Representation

Nobusumi Sagara†

†Faculty of Economics, Hosei University
4342, Aihara, Machida, Tokyo, 194-0298, Japan
Email: nsagara@mt.tama.hosei.ac.jp

Abstract—The purpose of this paper is twofold. First, we present the existence theorem of an optimal trajectory in a nonconvex variational problem with recursive integral functionals by employing the norm-topology of a weighted Sobolev space. We show the continuity of the integral functional and the compactness of the set of admissible trajectories. Second, we show that a recursive integrand is represented by a normal integrand under the conditions guaranteeing the existence of optimal trajectories. We also demonstrate that if the recursive integrand satisfies the convexity conditions, then the normal integrand is a convex function. These results are achieved by the application of the representation theorem in $L^p$-spaces.

1. Introduction

Let $\Omega$ be a bounded open subset of the real line and $f$ be a Carathéodory function on $\Omega \times \mathbb{R}^n \times \mathbb{R}^n$ satisfying the growth condition. It is well known that the integral functional $I$ on the Sobolev space $W^{1,p}(\Omega; \mathbb{R}^n)$ defined by $I(x) = \int_{\Omega} f(t, x(t), x'(t))dt$ is weakly lower semicontinuous if and only if $f(t, x, \cdot)$ is a convex function on $\mathbb{R}^n$ for any $(t, x) \in \Omega \times \mathbb{R}^n$ (see Dacorogna [7, Theorems 3.1 and 3.4]). The equivalence between the weak lower semicontinuity of integral functionals and the convexity of Carathéodory integrands has a long history and the variants of this result, dating back to Tonelli, have been extensively studied by various authors (see the references cited in Chapters 3 and 4 of Dacorogna [7]).

According to this equivalence, it is necessarily impossible for integral functionals to be weakly lower semicontinuous whenever Carathéodory integrands lack convexity, which easily leads to the nonexistence of an optimal trajectory in variational problems without convexity assumptions. This poses a serious problem especially when one treats a recursive integral functional because it entails a variable discount factor that depends on trajectories cumulatively in a nonexponential manner, and hence assuming the convexity of the integrand to guarantee the weak lower semicontinuity of the integral functional imposes a strong restriction on the variational problem. (For the economic motivation for introducing recursive integral functionals, see Becker et al. [2]). To evade this difficulty with nonconvex variational problems, one cannot help giving up the use of the weak topology of $W^{1,p}(\Omega; \mathbb{R}^n)$.

The purpose of this paper is twofold. First, we present the existence theorem of an optimal trajectory in a nonconvex variational problem with recursive integral functionals by employing the norm-topology of a weighted Sobolev space. The approach is a “direct method” of calculus of variations. We show the continuity of the integral functional and the compactness of the set of admissible trajectories. Existence follows from the classical Weierstrass theorem. This is a further generalization of the existence result by Chichilnisky [6] and Sagara [11] who worked with a Hilbert space with $L^2$-norm topology in that this paper considers more general recursive integral functionals and significantly weaker boundedness assumptions on the set of admissible trajectories than in the above works.

Second, we show that a recursive integrand is represented by a normal integrand under the conditions guaranteeing the existence of optimal trajectories. Therefore, the consideration of a recursive integral functional for the existence problem reduces to the consideration of a standard integral functional. We also demonstrate that if the recursive integrand satisfies convexity conditions, then the normal integrand is a convex function. These results are achieved by the application of the representation theorem in $L^p$-spaces obtained by Buttazzo and Dal Maso [3].

Use of the weighted Sobolev space with the norm-topology instead of the weak topology has two motivations. First, because strengthening a topology makes it easier for functions to be continuous, by considering the norm-topology, which is indeed stronger than the weak topology, we do not rely on the convexity of Carathéodory integrands in proving the continuity of integral functionals, so the continuity argument becomes relatively simplified. This is due to the significant result of Carathéodory functions and Nemitskii operators provided by Krasnosel’ski [9].

Second, because of the consideration of an unbounded interval $\Omega = [0, \infty)$ in this paper, we need a delicate treatment for the integrability of integral functionals. The trajectories under investigation may admit unboundedness under the norm of $W^{1,p}(\Omega; \mathbb{R}^n)$, but the growth rate of the trajectories is bounded by some weight function. The space of this type is described properly by a weighted Sobolev space, which is identified with a closed vector subspace of an $L^p$-space. The integrability of the recursive integral
functional is ensured on the set of admissible trajectories in the weighted Sobolev space under the growth condition on the Carathéodory integrand and the discount factor.

The weighted Sobolev space with the norm-topology was first introduced by Chichilinsky [6] into optimal growth theory to prove the existence of an optimal trajectory without convexity assumptions. Sagara [11] extended the existence result of Chichilinsky to the case of recursive integral functionals. The existence result for the weighted Sobolev space with the weak topology under the convexity assumptions was provided by [10]. The first rigorous treatment of the existence problem for the case of recursive integral functionals is that of Becker et al. [2], whose proof relies on the convexity of the set of admissible trajectories in the space of locally absolutely continuous functions with the weak topology and the convexity of the recursive integrand. Carlson [4] and Balder [1] extended the existence result of Becker et al. [2] for the optimal control problem with recursive integral functionals under convexity hypotheses by employing the lower closure theorem of Cesari [5].

2. Weighted Sobolev Space

Let \( \Omega = [0, \infty) \) be a half-open interval in \( \mathbb{R} \) and \( \mathcal{F} \) be the \( \sigma \)-field of Borel subsets in \( \Omega \). We denote by \( L^p(\Omega; \mathbb{R}^n) \) the set of \( \mathbb{R}^n \)-valued measurable functions \( u \) on \( \Omega \) satisfying \( ||u||_p = \left( \int_{\Omega} |u(t)|^p dt \right)^{1/p} < \infty \), where \( | \cdot | \) denotes the Euclidean norm of \( \mathbb{R}^n \) and the relevant integral is the Lebesgue integral on \( \Omega \). Let \( C_0^\infty(\Omega; \mathbb{R}^n) \) be the set of \( \mathbb{R}^n \)-valued smooth functions with compact support in \( \Omega \). If an element \( u \) in \( L^p(\Omega; \mathbb{R}^n) \) has its distributional derivative \( \mathcal{D}u \) in \( L^p(\Omega; \mathbb{R}^n) \), i.e., \( \int_{\Omega} \langle \mathcal{D}u(t), \varphi'(t) \rangle dt = -\int_{\Omega} \langle u(t), \varphi(t) \rangle dt \), for any \( \varphi \in C_0^\infty(\Omega; \mathbb{R}^n) \), where \( \langle \cdot, \cdot \rangle \) is the inner product of \( \mathbb{R}^n \), define the norm of \( u \) by \( ||u||_p = |||u||_p|^p + |||\mathcal{D}u||_p|^p \) for \( 1 \leq p < \infty \). The Sobolev space \( W^{1,p}(\Omega; \mathbb{R}^n) \) is the set of elements \( u \) in \( L^p(\Omega; \mathbb{R}^n) \) satisfying \( ||u||_p < \infty \).

Let \( \rho \) be a positive measurable function on \( \Omega \). The weighted \( L^p \)-space \( L^p_\rho(\Omega; \mathbb{R}^n) \) with a weight function \( \rho \) is the set of \( \mathbb{R}^n \)-valued measurable functions \( u \) on \( \Omega \) satisfying \( ||u||_{\rho,p} = \left( \int_{\Omega} |u(t)|^p \rho(t) dt \right)^{1/p} < \infty \). If an element \( u \) in \( L^p_\rho(\Omega; \mathbb{R}^n) \) has its distributional derivative \( \mathcal{D}u \) in \( L^p_\rho(\Omega; \mathbb{R}^n) \), define the norm of \( u \) by \( ||u||_{\rho,p} = |||u||_p|^p + |||\mathcal{D}u||_p|^p \) for \( 1 \leq p < \infty \). The weighted Sobolev space \( W^{1,p}_\rho(\Omega; \mathbb{R}^n) \) with a weight function \( \rho \) is the set of elements \( u \) in \( L^p_\rho(\Omega; \mathbb{R}^n) \) satisfying \( ||u||_{\rho,p} < \infty \). Under this norm \( W^{1,p}_\rho(\Omega; \mathbb{R}^n) \) is a separable Banach space. By the identification of each element \( u \) in \( W^{1,p}_\rho(\Omega; \mathbb{R}^n) \) with \( (u, u) \), it is obvious that \( W^{1,p}_\rho(\Omega; \mathbb{R}^n) \) is a closed vector subspace of the direct sum \( L^p_\rho(\Omega; \mathbb{R}^n) \oplus L^p_\rho(\Omega; \mathbb{R}^n) \).

3. Main Result

The variational problem under investigation consists of minimizing the recursive integral functional \( I : W^{1,p}_\rho(\Omega; \mathbb{R}^n) \to \mathbb{R} \cup \{ \infty \} \) having the form

\[
I(x) = \int_{\Omega} \left[ f(t, x(t), \dot{x}(t)) F(t, \int_0^t r(s, x(s), \dot{x}(s))ds) \right] dt,
\]

over the set of trajectories \( x \) in \( W^{1,p}_\rho(\Omega; \mathbb{R}^n) \) satisfying the differential inclusion

\[
\dot{x}(t) \in \Gamma(t, x(t)) \quad \text{a.e. } t \in \Omega \text{ and } x(0) \in X_0,
\]

where \( f \) and \( r \) are measurable functions on \( \Omega \times \mathbb{R}^n \times \mathbb{R}^n \), \( F \) is a measurable function on \( \Omega \times \mathbb{R} \) and \( \Gamma : \Omega \times X \to 2^{\mathbb{R}^n} \) is a set-valued mapping, and \( X \) and \( X_0 \) are subsets of \( \mathbb{R}^n \) with \( X_0 \subset X \). In most applications, \( t \in \Omega \) is time, \( f \) is a cost function, \( r \) is a discounting function and \( F \) is a variable discount factor in which the integral \( \int_0^t r(s, x(s), \dot{x}(s))ds \) takes into consideration of the cumulative dependence on past trajectories.

Define the set of admissible trajectories with initial conditions in \( X_0 \) by

\[
\mathcal{X}_\Gamma = \left\{ x \in W^{1,p}_\rho(\Omega; \mathbb{R}^n) \mid \dot{x}(t) \in \Gamma(t, x(t)) \text{ a.e. } t \in \Omega \text{ and } x(0) \in X_0 \right\}.
\]

Then the problem is to find a solution to

\[
\min \{ I(x) \mid x \in \mathcal{X}_\Gamma \}.
\]

(P)

3.1. Existence Theorem

The following assumptions guarantee the continuity of the recursive integral functional and the compactness of the set of admissible trajectories.

**Assumption 3.1.** (i) \( f(t, \cdot, \cdot) \) is continuous on \( \mathbb{R}^n \times \mathbb{R}^n \) a.e. \( t \in \Omega \) and \( f(\cdot, x, y) \) is measurable on \( \Omega \) for any \( (x, y) \in \mathbb{R}^n \times \mathbb{R}^n \).

(ii) There exist some \( \alpha \in L^1_\rho(\Omega; \mathbb{R}^n) \) and \( a_1, a_2 > 0 \) such that

\[
|f(t, x, y)| \leq \alpha(t) + a_1 |x|^p + a_2 |y|^p
\]

for any \((t, x, y) \in \Omega \times \mathbb{R}^n \times \mathbb{R}^n\).

(iii) \( F(t, \cdot) \) is continuous on \( \mathbb{R} \) a.e. \( t \in \Omega \) and \( F(\cdot, x, z) \) is measurable on \( \Omega \) for any \( z \in \mathbb{R} \).

(iv) \( r(t, \cdot, \cdot) \) is continuous on \( \mathbb{R}^n \times \mathbb{R}^n \) a.e. \( t \in \Omega \) and \( r(\cdot, x, y) \) is measurable on \( \Omega \) for any \((x, y) \in \mathbb{R}^n \times \mathbb{R}^n\).

(v) There exists some \( \beta \in L^1_{\infty}(\Omega) \) such that

\[
|r(t, x, y)| \leq \beta(t) \quad \text{a.e. } t \in \Omega \text{ for any } (x, y) \in \mathbb{R}^n \times \mathbb{R}^n
\]

and

\[
F(t, \int_0^t \beta(s)ds) \leq \rho(t) \quad \text{a.e. } t \in \Omega.
\]
Assumption 3.2. (i) $\mathcal{X}_t$ is closed in $W^{p}_Ω(\mathbb{R}^n)$.

(ii) There exists some $µ ∈ L^{p}(Ω, µ; R^n)$ such that $x ∈ X_\bar{t}$ implies $\max \{|x(t)|, |x(t)|\} \leq µ(t)$ a.e. $t ∈ Ω$.

Theorem 3.1. Suppose that the weight function $ρ$ is continuous. If Assumptions 3.1 and 3.2 are satisfied, then the problem (P) has a solution.

In general, it is ambiguous which requirements Assumption 3.2 imposes on $Γ$. The sufficient conditions on $Γ$, which imply Assumption 3.2, are provided by the following result.

Theorem 3.2. Let $A(t)$ be the t-section of $Γ$, that is,

\[ A(t) = \{(x, y) ∈ R^n \times R^n | y ∈ Γ(t, x)\}. \]

Suppose that the following conditions are satisfied:

(i) $X_0$ is closed in $R^n$.

(ii) $A(t)$ is closed a.e. $t ∈ Ω$ in $R^n × R^n$.

(iii) There exists some $µ ∈ L^{p}(Ω, µ; R^n)$ such that $x ∈ X_\bar{t}$ implies $\max \{|x|, |y|\} \leq µ(t)$ a.e. $t ∈ Ω$.

Then $\mathcal{X}_t$ satisfies Assumption 3.2.

Note that Theorem 3.1 is true for any choice of continuous weight functions. If one wishes to treat a weighted Sobolev space as broadly as possible, it is desirable to choose a weight function as small as possible because $0 < ρ_1 ≤ ρ_2$ implies that $W^{1,p}_ρ(Ω; R^n) ⊂ W^{1,p}_ρ(Ω; R^n)$.

3.2. Representation Theorem

For the representation theorem we require that the origin in $R^n$ belongs to the set of admissible trajectories and the value of the recursive integrand at the origin is normalized to zero.

Assumption 3.3. (i) $f(t, 0, 0)F(t, \int_0^t r(s, 0, 0)ds) = 0$ a.e. $t ∈ Ω$.

(ii) $0 ∈ Γ(t, 0)$ a.e. $t ∈ Ω$ and $0 ∈ X_0$.

A function $g : Ω × R^n × R^n → R ∪ \{+∞\}$ is a normal integrand if $g(t, ·, ·)$ is lower semicontinuous on $R^n × R^n$ a.e. $t ∈ Ω$ and $g(·, x, y)$ is measurable on $Ω$ for any $(x, y) ∈ R^n × R^n$.

The next result implies that for any primitive $(f, r, F, Γ)$ the recursive integral functional $I$ is represented by a normal integrand such that $f(x) = \int_0^\infty g(t, x(t), 0)p(t)dt$ on $\mathcal{X}_t$ under the same hypotheses guaranteeing the existence of an optimal trajectory.

Theorem 3.3. Suppose that the weight function $ρ$ is continuous. If Assumptions 3.1 to 3.4 are satisfied, then there exists a unique normal integrand $g : Ω × R^n × R^n → R ∪ \{+∞\}$ with the following properties:

(i) There exist some $a ∈ L^{1,ρ}_ρ(Ω)$ and $a_1, a_2 ≥ 0$ such that

\[ -\langle a(t) + a_1|x|^{\rho} + a_2|y|^{\rho} \rangle ≤ g(t, x, y) \text{ a.e. } t ∈ Ω \]

for any $(x, y) ∈ R^n × R^n$.

(ii) For any $x ∈ \mathcal{X}_t$ and $A ∈ F$:

\[
\int_A f(t, x(t), 0)F\left(t, \int_0^t r(s, x(s), 0)ds\right)dt = \int_A g(t, x(t), 0)p(t)dt.
\]

The following assumption provides the sufficient condition for the convexity of the recursive integrand $I$ on $W^{1,ρ}_ρ(Ω; R^n)$.

Assumption 3.4. (i) $f(t, x, y) ≥ 0$ a.e. $t ∈ Ω$ for any $(x, y) ∈ R^n × R^n$.

(ii) $F(t, z) ≥ 0$ a.e. $t ∈ Ω$ for any $z ∈ R$ and $F(t, ·)$ is increasing on $R$ a.e. $t ∈ Ω$.

(iii) $f(t, ·, ·)F(t, ·)$ is convex on $R^n × R^n$ a.e. $t ∈ Ω$.

(iv) $r(t, ·, ·)F(t, ·)$ is convex on $R^n × R^n$ a.e. $t ∈ Ω$.

Theorem 3.4. Suppose that the weight function $ρ$ is continuous. If Assumptions 3.1 to 3.4 are satisfied, then the normal integrand $g$ in Theorem 3.3 is a convex integrand, that is, $g(t, ·, ·)$ is convex on $R^n × R^n$ a.e. $t ∈ Ω$.

4. Proof of the Main Result

The proof of Theorem 3.1 is an immediate consequence of the classical Weierstrass theorem. To this end, it suffices to show that $\mathcal{X}_t$ is compact in $W^{1,ρ}_ρ(Ω; R^n)$ and $I$ is continuous on $\mathcal{X}_t$. For the compactness argument we employ the characterization of the norm compactness of $L^{1,ρ}_ρ(Ω; R^n)$. For the continuity argument, we need the continuity property of Nemitskii operators, which are defined by Carathéodory functions. General results on the continuity of Nemitskii operators are given by Krasnosel’skii [9]. The proof of Theorem 3.3 is based on the representation of nonlinear functionals on $L^p$-spaces by Carathéodory integrals provided by Buttazzo and Dal Maso [3].

4.1. Continuity of the Recursive Integral Functional

A function $f$ on $Ω × R^n$ is a Carathéodory function if $f(t, ·)$ is continuous on $R^n$ a.e. $t ∈ Ω$ and $f(·, u)$ is measurable on $Ω$ for any $u ∈ R^n$. It can be shown that Carathéodory functions are jointly measurable in $(t, u)$ on $Ω × R^n$. Let $f$ be a Carathéodory function on $Ω × R^n$. Nemitskii operator $T_t$ transforms each $R^n$-valued measurable function $u$ on $Ω$ into a measurable functions on $Ω$ by $(T_tu)(t) = f(t, u(t))$. Let $µ$ be a nonatomic measure of the Borel measurable space $(Ω, F)$. We denote by $L^p(Ω, µ; R^n)$ the set of $R^n$-valued measurable functions $u$ on $Ω$ satisfying $\int_Ω |u|^p dµ < ∞$. 

341
Proposition 4.1 (Krasnosel’skii, Nemytskii, Vainberg).

For every $1 \leq p < \infty$, it follows that:

(i) If a Nemytskii operator $T_f$ transforms every function in $L^p(\Omega, \mu; \mathbb{R}^n)$ into a function in $L^1(\Omega, \mu)$, then it is continuous and bounded.

(ii) A Nemytskii operator $T_f$ transforms every function in $L^p(\Omega, \mu; \mathbb{R}^n)$ into a function in $L^1(\Omega, \mu)$ if and only if there exist some $\alpha \in L^1(\Omega, \mu)$ and $a > 0$ such that

\[ |f(t, u)| \leq \alpha(t) + a|u|^p \quad \text{for any } (t, u) \in \Omega \times \mathbb{R}^n. \]

\[ \int_{\Omega} \int_{\mathbb{R}^n} f(t, u) \, d\mu(t) \, du = 0. \]

\[ \int_{\Omega} |u(t)|^p \, d\mu(t) = 0. \]

Theorem 4.1. $I$ is continuous on $\mathcal{D}_I$.

4.2. Compactness of the Set of Admissible Trajectories

The following result provides a useful criterion to check whether a subset of an $L^p$-space is norm compact. The proof is found in Dunford and Schwartz [8, Theorem IV.8.20].

Proposition 4.2. For every $1 \leq p < \infty$, a bounded subset $K$ of $L^p(\Omega; \mathbb{R}^n)$ is relatively compact if and only if the following conditions are satisfied:

(i) $\limsup_{\varepsilon \to 0} \int_{\Omega} |u(t + \varepsilon) - u(t)|^p \, dt = 0.$

(ii) $\limsup_{t \to \infty} \int_{\Omega} |u(t)|^p \, dt = 0.$

Theorem 4.2. $\mathcal{D}_I$ is compact in $W^{1, p}_\mu(\Omega; \mathbb{R}^n)$.

4.3. Representation of Recursive Integral Functionals by Normal Integrands

The following general result on the representation of functionals on $L^p$-spaces is due to Buttazzo and Dal Maso [3].

Proposition 4.3 (Buttazzo and Dal Maso). Let $1 \leq p < \infty$ and $G : L^p(\Omega, \mu; \mathbb{R}^n) \times \mathcal{F} \to \mathbb{R} \cup \{+\infty\}$ be a functional satisfying the following conditions:

(i) $G(\cdot, \Omega)$ is lower semicontinuous.

(ii) There exists some $u_0 \in L^p(\Omega, \mu; \mathbb{R}^n)$ such that $G(u_0, A) < +\infty$ for any $A \in \mathcal{F}$.

(iii) $G$ is local on $\mathcal{F}$, that is, $u, v \in L^p(\Omega, \mu; \mathbb{R}^n)$ and $u = v \mu$-a.e. on $A \in \mathcal{F}$ imply $G(u, A) = G(v, A)$.

(iv) $G$ is finitely additive on $\mathcal{F}$, that is, $u \in L^p(\Omega, \mu; \mathbb{R}^n)$, $A, B \in \mathcal{F}$ and $A \cap B = \emptyset$ imply $G(u, A \cup B) = G(u, A) + G(u, B)$.

Then there exists a unique normal integrand $g : \Omega \times \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ with the following properties:

(1) There exist some $\alpha \in L^1(\Omega, \mu)$ and $a \geq 0$ such that

\[ -(\alpha(t) + a|x|^p) \leq g(t, x) \quad \text{for any } x \in \mathbb{R}^n. \]

Moreover, if $G(\cdot, \Omega)$ is weakly lower semicontinuous, then $g$ is a convex integrand, that is, $g(t, \cdot)$ is convex $\mu$-a.e. $t \in \Omega$ on $\mathbb{R}^n$.

References


Analysis of Artificial Double Auction Markets Consisting of Multi-agents using Genetic Programming for Learning and its Applications to Control

Yoshikazu IKEDA† and Shozo TOKINAGA‡

†Faculty of Economics, Shinshu University
3-1-1 Asahi, Matsumoto, Nagano 390-8621 JAPAN
‡Graduate School of Economics, Kyushu University
6-19-1 Hakozaki, Higasi-ku, Fukuoka-shi, 812-8581 JAPAN

Abstract—In this paper, we show the analysis of artificial double auction markets consisting of multi-agents who learn based on the Genetic Programming (GP) and its applications to control. Agents in the market are allowed to buy or sell items depending on the prediction of situations. In simulation studies, the change of bid price and the return of bidders is discussed depending on the demand curve of customers. We also propose a stabilization scheme of bid prices in double auction markets.

1. Introduction

Auction mechanisms are now expected to bring deregulation and competitive prices in markets of items such as network access rights, telecommunication channels, gas and electric utilities[1]. However, most of theoretical researches in auction theory assume that bidders will likely be competitive, and these bidders are symmetrical in size and are risk neutral, and bid only a unique good.

In this paper, we show the analysis of artificial double auction markets consisting of multi-agents who learn from past experiences based on the GP and its applications to control. In previous works, we demonstrated that in the double auction markets consisting multi-agents using GP learning, the bid prices bear very large changes if the demand of goods is not stationary [2]. But, the meaningful result for the control of price changes is not given.

We apply the GP procedure to model learning of agents[5]-[8]. Each agent has a pool of individuals represented in tree structures to decide future bid price and the volume of items to be supplied to the market using the past result of auctions. The fitness of individuals is defined by using the successful bids and the utilization of units, and agents improve their individuals based on the fitness to get higher return in coming auctions.

In the simulation studies, we can see the changes of prices, relation between and the structure of bidding and the profit of bidders. Moreover, we apply approximate method for the control of chaos to stabilize the changes in bid prices.

2. System configuration

2.1. Characteristics of auction market

We consider $N$ agents who behaves as an electric power producer with one generating unit and wishes to schedule its unit to maximize profit over a short time period of length $T$ hours[3][4]. Even though the description of problems is restricted to the electricity market, we can easily extend the frameworks to another production systems of goods.

The evaluation function for optimal generation of electricity includes following variables[3][4].

- $P_t$:production amount of power at $t$
- $CF(p)$:cost function of unit to generate power $p$
- $l_t$:volume to be sold under contracts
- $R$:price of electricity to be sold
- $E_t$:amount to be bought (if $E_t$ is negative, to be sold) from auction market
- $m_t$:market price at power pool

The objective function to maximize expected total profit is given by

$$\max \ E\left( \sum_{t=1}^{T} PR_t \right), \quad PR_t = l_t R - m_t E_t - CF(P_t).$$

under the constraints

$$E_t + P_t = l_t, \quad P_t = \lambda_t P_c,$$

where $P_c$ is the highest (full) capability to supply electricity, and the coefficient $\lambda_t$ denotes the utilization of unit. The amount $P_t$ of generation of electricity is given through $P_c$ multiplied by the utilization $\lambda_t$. If $E_t$ is positive, the electricity is bought from the pool, otherwise sold. Assuming that the price $m_t$ and $P_t$ are estimated based on the GP methods.

The generation of electricity $P_t$ is defined by the Cobb-Douglas type production function as follows.

$$P_t = A_t V_t^\alpha_t K_t^\beta_t.$$
of generation of electricity. For simplicity, we assume these efficiency parameters are time-invariant all through the time, and then we drop the subscript t for the parameters. We also assume that $\alpha + \beta < 1$. Assuming that the unit cost for the variable factor $V_i$ is $w_i$, and the unit cost for the fixed factor is $r_i$, then the limit cost function $CF(P_t)$ is given by

$$CF(P_t) = w_i V_i + r_i K_i.$$  

If we assume that the fixed factor is constant through the time ($K_t = K$), then we have the marginal cost function $MC_t$ as follows.

$$MC_t = \alpha^{-\frac{1}{2}} K^{-\frac{\beta}{\alpha}} w_i P_t^{1+\frac{1}{2}}.$$  

In the paper, we use the model that agents improve determination functions for the market price of electricity $m_t$ and the amount of relevant purchase of electricity $E_t$ (in cases of $E_t$ is negative, then they supply the electricity) based on the GP procedure. Agents exhibit these prices and amount for purchase (supply) to the market for bids in the double auctions. Then, the relevant level of generation of electricity $P_t$ for their own units are determined.

3. Market price and agents’ behavior

In the multi-unit auction, bidders make sealed bids indicating the quantity of goods they are willing to buy at a range of prices. The auctioneer allocate the goods to the highest bid first, and so on down the sequence of received bids until all the goods have been allocated. Likewise, in uniform price auctions successful bidders all pay the same price regardless of the bids they actually made. The price is equal to the highest (marginal) bid price accepted (called Pay Marginal).

Agents utilize the GP procedure to decide the market price $m_t$ based on past experiences. Then, agents can decide optimal value of generation $P_t$ and utilization of units $\lambda_t$.

As a result, agents appear as sellers in the market as well as bidders, since they buy the electricity if they think that the cost to generate electricity is costly than buying electricity. However, the producers are originally join to the market as suppliers using their facilities, then the utilization of units affect the return on initial investments. Then, we introduce the utilization of units as another measure to define the fitness of individuals.

It is assumed that the first $N_1$ times of biddings are used for learning for agents, and no item is delivered to bidder, and sellers get no money. In this learning period, each agent try to improve the estimation of individuals by using the GP procedure. Then, in successive $N_2$ times of bidding, agents apply the estimation using the pool of individuals. After $N_2$ times of auctions, the profit of each agent is determined.

4. GP learning and behavior of agents

In sealed-bid auctions, agent $i$ define the private evaluation $v_i$ denoting the preference of bid price. It is that agents are satisfied if the bid price in the auction is lower than the private evaluation. In the simulation studies, we define the private evaluations as the price of electricity $R$ under contracts between producers and customers.

We assume three cases for the definition of private evaluations $v_i$ as follows.

(Case 1) identical
(Case 2) uniformly distributed
(Case 3) piecewise constant

In the following, we assume that agents learn from past results of auctions to find appropriate bid prices and volumes of items for future auction based on the GP. In the GP, the system equations are represented in the tree structure (called individuals). In the parse tree, non-terminal nodes are taken from some well-defined functions such as binomial operation $+, -, \times, /$, and the operation taking the square root of variable. The equation represented by using the prefix is called as an individual. By using the measure of fitness to evaluate each individual, we apply the GP to the population to derive better description for future auctions promising higher profit.

The function determining bid price at time $t$ can have also the symbols $P_1=CP(t-1)$ where $CP(t-1)$ is the successful bid price in previous auction. The function has also symbols $\land, \lor, \max$ and $\min$ defined by taking the average, maximum and minimum of successful bid prices in previous $t_1, t_2, t_3$ time periods of auctions, respectively.

We iteratively perform the following steps until the termination criterion has been satisfied[5]-[10].

(Step 1) Generate an initial population of random composition of possible functions and terminals for the problem at hand. The random tree must be syntactically correct program.

(Step 2) Execute each individual (evaluation of system equation) in population, then, assign it a fitness value giving partial credit for getting close to the correct output.

(Step 3) Select a pair of individuals chosen with a probability $p$, based on the fitness.

(Step 4) Then, create new individuals (offsprings) from the selected pair by genetically recombining randomly chosen parts of two existing individuals using the crossover operation applied at a randomly chosen crossover point.

(Step 5) At a certain probability, we apply the mutation operation for the pool of individuals. If the result designation is obtained by the GP (the maximum value of the fitness become larger than the prescribed value), then terminate the algorithm, otherwise go to Step 2.

The ability of individuals corresponding to the functions
is defined as the fitness in the GP. As the first ability measure, we use following value based on the number of successful bids.

\[ pr_{ik} = \sum_j PR_j / N_{ik}. \]  

(7)

where \( N_{ik} \) is the number of successful bid by the agent \( i \) obtained by using \( k \)-th individual in the pool. We also employ the utilization rate \( \lambda_i \) as the second evaluation measure for fitness in \( k \)-th individual for \( i \)-th agent.

Finally, by changing the weight \( \omega_i \) between \( pr_{ik} \) and \( \lambda_i \), we have aggregated fitness measure for \( k \)-th individual as follows.

\[ s_{ik} = \omega_i (pr_{ik} - \min_j pr_{ij}) / R^{pr}_t + (1 - \omega_i) \lambda_i. \]  

(8)

where \( R^{pr}_t \) is the range of profit \( pr_{ik} \) in equation (8) so that we can normalize the profit between 0 and 1 by subtracting the minimum value form the profits and by dividing them by the range to make two terms in the fitness functions to be comparable.

5. Applications

The parameters for simulation studies are selected as follows.

- Number of agents: 20
- \( N_1 = 500000 \) (apply GP operation for each 1000 auctions)
- \( N_2 = 500000 \) (upper limit of bid price: 150)
- Number of individuals for each pool: 50
- Maximum number of nodes in trees: 50
- Probability of crossover: 0.05
- Probability of each mutation: 0.05
- Parameters: \( \alpha = 0.1, \beta = 0.4, \omega_i = 1, \lambda_i = 100 \)

In terms of the parameters \( A, K \) defining the efficiency of units are given for two groups \( A \) and \( B \) each of which includes half of agents. For the group \( A(B) \), we assign \( A = 1, K = 10 \) (\( A = 5, K = 100 \)), and that means agents in the group \( A(B) \) have units with relatively lower (higher) efficiency. Then, in principle, agents in Group \( A(B) \) tend to be buyers (sellers) of electricity.

Fixed demand

At first, we consider the case where the demand of customers is fixed (constant). Figure 1 shows an example of time series of bid price obtained after sufficient time of biddings. As is seen from the figure, even though the bid price is stable in the range from 90 to 100, but the bid price is still fluctuating and does not converge to a certain constant level. The contract price of electricity is fixed to be 100, but the price of electricity which is actually traded in the market is 101.4 in average. The fact means that in the double auction market, the trades are basically preferable for sellers than for buyers, and buyers are forced to purchase the goods at relatively higher prices than realizable prices.

In this case, the average profit of agents in Group \( A(B) \) is -6252 (4736), and that mean agents in Group \( A \) are forced to generate at higher cost than the contracted price, and

![Figure 1: Example of time series of auction price](image)

even more they need to purchase electricity from the market at higher prices. On the other hand, agents in Group \( B \) possessing units with higher efficiency can get higher profit by selling electricity.

Time-varying demands

Then, we examine the cases where the demands of customers are time-varying and are not constant along the time. The time series of the electricity demand is assumed to be a random walk which is obtained by adding an incremental value selected from \(-2, -1, 0, 1, 2\) at the same probability to the previous demand value. For convenience, the range of demand is restricted between 50 and 150.

Figure 2 shows an example of time-varying demands of electric power. And figure 3 shows an example of the time series of bid price under time-varying demand of customers. As is seen from Figure 3, the bid price bears changes and fluctuations, and sometime the ranges between the highest and lowest price are four or five times larger than the range of prices for time-invariant demands of customers. Moreover, we find impulsive (sudden) rise of bid price having about ten times larger amplitude than the price for time-invariant demand cases. The facts implies us that the offers of item done by sellers induce the reaction of bidders, and as a result, a small change of bid price is enlarged (exaggerate) in the bidding process, and the jumps in bid price are observed.

6. Control scheme

Consider a non-linear dynamic system

\[ x(t+1) = f(x(t), u(t)). \]  

(9)

where we assume that \( x(t) \) is the state variables, and \( u(t) \) is the only available control parameter which we allow to vary in a range. We assume further that we are not far apart from the neighborhood of some steady state \( x_f \) (fixed point), which we want to stabilize by choosing an appropriate sequence of admissible input.

We assume that the dynamic system \( f(x(t)) \) with \( u(t) = 0 \) is estimated by using the GP, and is denoted as \( f(x(t)) \).
Then, the control method is derived straightforward by using the approximation. Since we can obtain the estimated value \( \hat{x}(t + 1) \) for the next state, we impose the input \( u(t) \) so that

\[
x_f = \hat{x}(t + 1)' = \hat{f}(x(t), u(t)).
\]

The optimal value \( u^*(t) \) of input \( u(t) \) is estimated by incrementally changing \( u(t) \).

In the electric power auction, \( x_f \) is the target price of the auction and \( u(t) \) is a bidding price and quantity (2 dimensional vector) of the electric power producer who has influence on price determination (price maker). Price maker predicts the price of the next auction by another GP at the same time it determines a bidding price and quantity by GP.

Then, the following value means the rate of convergence to the targeted level of price after the control which depends on \( x_f \).

\[
R(x_f) = \sum_{t=T_1+1}^{T_2} [CP(t) - x_f]^2/(T_2 - T_1).
\]

where \( CP(t) \) is contract price obtained by a simulation. \( T_1 \) is start time of the control, and \( T_2 \) is end time of it. (Simulation results are omitted here).

References


Abstract—There has been an argument whether foreign exchange rates are the random walk process or not. If not, then there is an other possibility that foreign exchange rates can have fractal nature. In this Paper we have tried to estimate the fractal nature of euro foreign exchange rates using the rescaled range(R/S) analysis, empirically compared with the random walk process. The results are relevant to our hypothesis. Further we have tried to find the number of periodic cycles within the time series, using V-Statistic. We discuss this matter using Weierstrass function, which is the first fractal function, continuous everywhere, but nowhere differentiable. This function is an infinite sum of a series of sine( or cosine) waves, in which the amplitude decreases, while the frequency increases according to different factors. It will be showed that R/S analysis is applicable to detect the number of periodic cycles.

1. What is Fractal

1.1 Charactaristic of Fractal

The term “Fractal” has no precise definition, but has several characteristics. The first property is the self-similarity. It means that a figure or time series data can be seen as similar, even if the scale is variously changed. And this characteristic is also related to the scale-invariant nature of fractal, which is explained by the power-law relation of variables.

1.2. Fractal Nature of Time Series Data

As for the fractal analysis of time series data, the fractional Brownian motion is of importance. Einstein has figured out more precisely the Brownian motion in his work (the erratic path followed by a particle suspendes in a fluid) in 1908. Brownian motion is a primary model for a random walk process. Here the distance that a random particle covers will increase with the square root of time. The equation used to measure it, is written as follows;

\[ R = T^{0.5} \]  (1.1)

where \( R \) = the distance covered
\( T \) = a time index.

This is called the “\( T \) to the one-half rule”, and is also the typical characteristic of a random walk process.

2. R/S Analysis

2.1. How We Can Calculate the Rescaled Range

Hurst(1951) has founded that an underlying periodic behavior could be detected with R/S analysis (rescaled range analysis).

In this chapter, we explain the method to calculate the R/S value, starting with a time series, \( x = x_1, x_2, \cdots, x_n \), to represent \( n \) consecutive values. The mean value \( x_m \) of the time series \( x \) is defined as:

\[ x_n = (x_1 + x_2 + \cdots + x_n)/n \]  (2.1)

The standard deviation, \( s_n \), is estimated as:

\[ s_n = n^{-1/2} \sqrt{\frac{1}{n}(x_r - x_m)^2} \]  (2.2)

which is merely the standard normal formula for standard deviation. The rescaled range was calculated by first rescaling or normalizing the data by subtracting the sample mean:

\[ Z_r = (x_r - x_m); \quad r = 1, \ldots, n \]  (2.3)

This series, \( Z_r \), has a zero mean. And the next step creates a cumulative time series \( Y \):

\[ Y_r = (Z_1 + Z_r) \quad r = 2, \ldots, n \]  (2.4)

The last value of \( Y_r \) will always be zero because \( Z \) has a mean of zero. Here we have the adjusted range, which is calculated by the maximum minus the minimum value of the \( Y_r \):

\[ R_n = \max(Y_1, \ldots, Y_n) - \min(Y_1, \ldots, Y_n) \]  (2.5)

This adjusted range, \( R_n \), is the distance that the system travels for time index \( n \). If we set \( n = T \), we can apply equation (1.1), provided that the time series, \( x \), is
independent for increasing values of $n$. But equation (1.1) is only limited to time series that are in Brownian motion; that is, they have zero mean, and variance equal to one.

To apply this concept to time series that are not in Brownian motion, there is a need to generalize equation (1.1), taking into account systems that are not independent. Hurst found that the following equation was a more generalized form of equation (1.1):

$$R/S = c \cdot n^H$$

where $C$ is a constant, and $H$ is called the Hurst exponent. R/S value of equation (2.6) is referred to as the rescaled range because it has zero mean and is expressed in terms of local standard deviation. This R/S value scales, as we have mentioned, increase the time increment, $n$, by a power-law value equal to $H$ exponent. Here we can say that the rescaled range analysis is a fractal analysis, because all fractal scale according to a power law. In a fractal time series, the range increases according to a power.

2.2. The Empirical Model

The Hurst exponent can be approximated by plotting the log (R/S) versus the log(n) and estimating the slope through an ordinary least squares regression. The econometric model for this estimation is quite simple as follows:

$$\log(R/S) = \log(c) + H \cdot \log(n)$$

If a system were independently distributed, then $H=0.5$. Hurst investigated the Nile River, and found that $H$ was 0.9. Thus the rescaled range of height of the river was increasing at a faster rate than the square root of time.

3. Empirical Analysis

3.1. Random Walk Simulation

As we already mentioned above, if the system were under a random walk process, then the Hurst exponent would be 0.5. To compare the empirical analysis of exchange rates with the theoretical value, we have tried to apply the R/S analysis to the system that is created by random walk simulation on the computer.

Figure 1 shows the log-log plot of one of random walk simulation, and the results of the simulation is described in table 1 on the last page of this paper.

3.2. R/S Value of Foreign Exchange Rates

We have applied R/S analysis to the daily euro exchange rates from the start day of this European single currency to February 2006. The data are reference rate of the European Central Bank, available from statistic section of the German Bundesbank.
Figure 2&3 shows the log-log plot of the R/S and n for the euro exchange rates against US dollar and Yen. The Hurst exponent of euro/yen exchange rates is estimated to 0.617, that is higher than the theoretical random walk of H=0.5. In the same way, the Hurst exponent of euro/dollar is estimated to 0.651, that is quite similar to the formar one. In this paper, we estimated R/S value of Pound Sterling, Swiss Franc, Swedish Krona, Polish Zloty, Hungary Forint, and Hong Kong Dollar additionally for the purpose of comparison.

4. Detecting Periodic Cycles in Time Series Data

4.1. Weierstrass Function

In the case of simple sine waves, the time series would be a bounded set. The R/S values would reach a maximum value after one cycles, because the range could not grow beyond the amplitude.

Karl Weierstrass, a german mathematician presented the historically first fractal function, which has composed of a infinite sum of a series of sine or cosine wave, in which the amplitude decreases, while the frequency increase according to different parameters. This function is nowhere differentiable. West(1990) has used this function as an analytical object for fractal time series.

The Weierstrass function superimposes an infinite number of sine waves. The equation of this function is written as a Fourie series as follows;

$$F(t) = \sum_{n=0}^{\infty} \frac{1}{a^n} \cos(b^n \cdot \omega \cdot t)$$  \hspace{1cm} (4.1)

We begin with a frequency, \(\omega\), with a amplitude of 1. A second harmonic term is summed, with frequency \(b\omega\) and amplitude \(1/a\), with \(a\) and \(b\) greater than 1. The third harmonic term has frequency \(b^2\omega\) and amplitude \(1/a^2\). The fourth term has frequency \(b^3\omega\) and amplitude \(1/a^3\). This progression continues indefinitely. Each term has frequency that is a power of \(b\) greater than the previous one, and amplitudes that is a power of a smaller.

Figure 4 shows a Weierstrass function using the first four terms ( \(n=1 \) to \(4\)).

4.2. V-Statistic

If we look at the log-log plot of exchange rates, the slope is not constantly flat or smooth, rather we can see “breaks” on the slope. V-statistic is of great use to see the breaks in the log-log plot, and to make a correct estimation of the cycle length, which functions well in the presence of noise. This V-statistic is defined as follows:

$$V_n = (R/S)_n / \sqrt{n}$$  \hspace{1cm} (4.2)

The graph of this value would have a horizontal line, if the R/S statistic was scaling according to the square root of time. That is; if the process was an independent, random walk, the plot of V versus log(n) would be flat.

On the other hand, if the process would be persistent, and R/S was scaling at a faster rate than the square root of time (H>0.5), then the line would be upwardly sloping. If the process was anti-persistent(H<0.5), then the graph would be downward sloping. The breaks would occur, when the V chart flattens out, in the point the long-memory vanishes. Figure 5 shows the V statistic for the Weierstrass function simulated, where we can see flattening parts about three times, which are seen like the peaks of each mountain.

Figure 5:

![V-Statistic: Weierstrass Function](image)

Figure 6:

![V-Statistic: euro/dollar FX-rate](image)
This flattening of the slope appears at the end of each periodic cycle. Therefore we can say that the R/S analysis is capable of determining periodic cycles in time series data. In other words, the R/S analysis can discern cycles within a cycle. For example, there can be at least 3 cycles within the time series of the simulated Weierstrass function.

Next, figure 6 shows the V statistic for the euro/dollar exchange rates. We can see a break point, or a peak, where the log(n) is about 2.3. As the result of this analysis, it is possible that the euro exchange rates are not random process, but have the persistent trend within a cycle, where each cycle length is to be estimated 200 dairy data.

5. Concluding Remarks

We have empirically estimated the R/S statistic of the euro exchange rates, and results are described in the table 1. The Hurst exponents of euro exchange rates are greater than 0.5, so that the euro exchange rates can be seen as a fractal time series.

The relation of the “box counting dimension” and fractal dimension $D_f$ is well-known described as follows:

$$D_f = \lim_{k \to \infty} \frac{\log(2^{2-H})^k N}{\log(2^k/r)} = 2 - H \quad (5.1)$$

where $r$ is a length of each box line, $N$ is the number of boxes.

As a result, the fractal dimension of euro/dollar exchange rates is calculated 1.35 by this method.

Second, there is an exception. Only the Hurst exponent of Hungary Forint was slightly under 0.5. We can say that this could be caused by the fact that Forint currency is under the fixed exchange rate system to euro, and therefore the range of the rates are quite limited in a narrow band.

Third important result is the fact that the R/S value of Hong Kong dollar is quite similar to that of US dollar. That could be occurred, perhaps because HK dollar is pegged in US dollar, therefore the behavior of this two currency against euro could be similar.

Table 1: Results of R/S analysis

<table>
<thead>
<tr>
<th>Currency</th>
<th>Hurst exponent</th>
<th>R squared</th>
<th>Fractal Dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yen</td>
<td>0.617</td>
<td>0.957</td>
<td>1.383</td>
</tr>
<tr>
<td>US Dollar</td>
<td>0.651</td>
<td>0.949</td>
<td>1.349</td>
</tr>
<tr>
<td>Pound Sterling</td>
<td>0.595</td>
<td>0.981</td>
<td>1.407</td>
</tr>
<tr>
<td>Swiss Franc</td>
<td>0.526</td>
<td>0.971</td>
<td>1.474 stable to euro</td>
</tr>
<tr>
<td>Swedish Krona</td>
<td>0.65</td>
<td>0.967</td>
<td>1.35</td>
</tr>
<tr>
<td>Poland Zloty</td>
<td>0.599</td>
<td>0.987</td>
<td>1.401</td>
</tr>
<tr>
<td>Hungary Forint</td>
<td>0.477</td>
<td>0.972</td>
<td>1.523 pegged to Euro</td>
</tr>
<tr>
<td>HK Dollar</td>
<td>0.647</td>
<td>0.945</td>
<td>1.353 pegged to follar</td>
</tr>
</tbody>
</table>

random walk 1 0.528 0.78 1.642
random walk 2 0.469 0.933 1.531
random walk 3 0.502 0.979 1.498

*Sample period: 4 Jan 1999-27 Feb 2006

References

Detecting nonlinearity in non-stationary time series

Tomomichi Nakamura\textsuperscript{a}, Michael Small\textsuperscript{a}, Yoshito Hirata\textsuperscript{b}

\textsuperscript{a} Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
Email: entomo@eie.polyu.edu.hk, ensmall@polyu.edu.hk
\textsuperscript{b} Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
Email: yoshito@sat.t.u-tokyo.ac.jp

Abstract—We describe a method for investigating nonlinearity in irregular fluctuations of time series, even if the data exhibit long term trends. Such situations are theoretically incompatible with the assumption of previously proposed methods. The null hypothesis addressed by our algorithm is that irregular fluctuations are generated by a stationary linear system. The method is demonstrated for numerical data generated by known systems and applied to several experimental time series.

1. Introduction

To investigate nonlinearity in irregular fluctuations various surrogate data methods have been proposed: the Fourier transform (FT), the amplitude adjusted Fourier transform (AAFT), and the iterative AAFT (IAAFT) algorithms [1, 3]. All of these techniques are linear surrogate methods [10], because they are based on a linear process and address a linear null hypothesis. These methods assume that data is stationary like Fig. 1(a). Unfortunately nonstationary data like Figs. 1(b-d) are theoretically incompatible with the assumption of linear surrogate tests and the nonstationarity is therefore very likely to lead to incorrect results [1, 3]. Dealing with such nonstationary data is difficult. In this paper, we introduce a method to investigate whether there is nonlinearity in irregular fluctuations (short term variability) even if they exhibit long term trends. The method is an intuitive modification of previously proposed linear surrogate methods.

2. Current technologies

The previously proposed methods are designed to generate flawless linear data [1, 3]. The basic strategy of these methods is as follows. One first applies the Fourier transform to the original data, randomizes the phases, and then inverts the transform using the randomized phases. In these methods all phases are randomized so as to eliminate any nonlinearity in data and then the data can be treated as completely linear\textsuperscript{1}. Although the methods are effective for irregular fluctuations, the methods are not effective for data

\textsuperscript{1}AAFT and IAAFT surrogate data are obtained by rank-ordering of the original data. Hence, in the most strict sense, these surrogate data may still include some kind of static nonlinearity.

Figure 1: Segments of four time series examined in this paper: (a) nuclear magnetic resonance (NMR) laser data, (b) monthly global average temperature (MGAT) from September 1920 to December 2005, (c) monthly sunspot numbers (MSN) from January 1749 to 10 August 2004, and (d) $x$ component of the Ikeda map data with artificial trends, which exhibit long term trends because the methods cannot preserve such trends.

3. Our proposed algorithm

To investigate nonlinearity in irregular fluctuations (especially when they are modulated by long term trends or periodicities), we want to destroy nonlinearity in irregular fluctuations and preserve the global behaviors. When data exhibit irregular fluctuations and long term trends the power spectrum is usually like Fig. 2. Figure 2 indicates that the data have large peaks of power in lower frequency domain and power in higher frequency domain is almost white. From this figure we conclude that the higher frequency domain is probably dominated by irregular fluctuations. This implies that even if we randomize phases in the higher frequency domain $f_e$ (see Fig. 2), the influence for long term trends will not be significant. Hence, we randomize phases only in the higher frequency domain and do not alter low frequency phases. In this way, long term trends are preserved in these unaltered low frequencies. This approach is in contrast to previously proposed linear surrogate methods, where all phases are randomized. We call our method “truncated Fourier transform surro-
Figure 2: The estimated power spectrum of the artificial data shown in Fig. 1(d), where we use 4096 data points. Note the logarithmic scale. We randomize phases in higher frequency domain \( f_e \) and other phases are untouched. The parameter \( f_e \) is the ratio of high frequency domain to the whole frequency domain. For example, when phases with frequency between 1500 and 2000 are randomized (that is, 500 higher frequency domain), \( f_e \) is 500/2000, that is, \( f_e = 0.25 \). We note that when showing power spectrum, these usually correspond to each frequency with unit of hertz (Hz) on the horizontal axis. In this paper, to explain our proposed method more easily we use arbitrary scale which correspond the number of data points.

3.1. Frequency domain to randomize phases

Obviously, the surrogate data generated by our method are influenced primarily by the choice of frequency domain \( f_e \) (see Fig. 2). The either too narrow or too wide domain is likely to lead to wrong judgement. However, we usually cannot determine an adequate value for \( f_e \) a priori. Hence, we increase \( f_e \) to randomize the phases from higher domain to lower domain step by step, for example by every 0.05 or 0.1. We continue until linearity and long term trends are preserved in the surrogate data. We describe the stopping criterion in detail in Sec. 6.

It should be noted that there is a possibility that only when all phases are randomized, nonlinearity in irregular fluctuations can be detected. Our algorithm clearly fails to detect the nonlinearity for such data. Hence, even if the NH is not rejected, we still cannot get past the possibility that the irregular fluctuations include nonlinearity. However, if the NH is rejected, it will be a strong evidence that there is some kind of nonlinearity in the irregular fluctuations.

4. The Fourier transform problem

There is a problem, wraparound effect, when there is a large difference between the first and last points. This effect introduces significant bias in the estimated linear properties of the power spectrum [5, 10]. Thus, if we use surrogate data generated in this way, we may wrongly judge the existence of nonlinearity in irregular fluctuations. More details concerning the relevant problems [3, 5, 10].

To ameliorate this artifact, when we calculate the power spectrum of such data, we symmetrize the original data first. By this procedure, there is no end point mismatch in the data. We find that the Fourier transform process is then not critically affected by the wraparound effect. Other operations are the same as the TFTS method. Hence, we call the method the symmetrized TFTS (STFTS) method. We apply the STFTS method to data with no trend and the STFTS method to data with long term trends.

5. How to reject a null hypothesis

Discriminating statistics are necessary for hypothesis testing. After calculation of the statistic, we need to inspect whether the NH shall be rejected or not.

We choose to use the average mutual information (AMI) as a discriminating statistic [7]. We consider different data realizations from the same population to have the same information flow so that the behavior of the AMI is the same.

To inspect whether a NH shall be rejected or not we employ Monte Carlo hypothesis testing. We check whether an estimated statistic of the original data falls within or outside the distribution of the surrogate data [2]. When the statistics fall within the distribution of the surrogate data, the NH may not be rejected. We generate 99 surrogate data and hence the significance level is 0.01 for a one-sided test.

6. Stopping criterion for increasing frequency domain

We need to inspect whether linearity and long term trends are preserved in surrogate data. We usually cannot know the adequate size of \( f_e \) over which to randomize phases. To know a rough upper bound of \( f_e \), we estimate the power spectrum of the original data. Then we determine the frequency domain where the power spectrum is almost white. This gives a good indication of the rough upper bound of the frequency domain to randomize.

Also, as mentioned above, we increase \( f_e \) to randomize the phases from higher domain to lower domain step by step. As \( f_e \) increases, we need to inspect whether linearity and long term trends are preserved in the surrogate data, although this is not necessary when the data have no long term trend. In addition to visual inspection, we inspect the

\[ f_e = \frac{500}{2000} = 0.25 \]
auto-correlation (AC) of the original data and the surrogate data at time lag 1 because the AC at time lag 1 must be most sensitive to the nature of the data. When the AC falls within the distribution, we consider that linearity and long term trends are sufficiently preserved in the surrogate data, and then calculate the AMI. When the AC falls outside the distribution, we do not use the data, stop increasing \( f_\varepsilon \) and adopt the last result.

7. Numerical Examples

We now demonstrate the application of our algorithm and confirm our theoretical arguments with several cases. As irregular fluctuations we consider: a linear autoregressive (AR) model and the Ikeda map. In all cases we use 4096 data points, and the data are both noise free and contaminated by 10% Gaussian observational noise. Irregular fluctuations are generated by following two models:

- The linear AR model given by \( x_t = a_1 x_{t-1} + a_6 x_{t-6} + \eta_t \) [8], where \( a_1 = 0.3, a_6 = 0.2 \) and \( \eta \) is Gaussian dynamical noise with standard deviation 1.0.
- The Ikeda map given by

\[
\begin{align*}
  f(x, y) &= \left( 1 + \mu (x \cos \theta - y \sin \theta), \mu (x \sin \theta + y \cos \theta) \right), \\
  \theta &= a - b \left( 1 + x^2 + y^2 \right), \\
  \mu &= 0.83, \quad a = 0.4, \\
  b &= 6.0 \quad [9].
\end{align*}
\]

These data are added to the artificial trends, and the level of additional data to the trends is equivalent to 10% (20dB) observational noise at each case. See the behaviors in Fig. 1(d). We apply the STFTS method to these data as an extreme case to demonstrate our premise.5

When irregular fluctuations are the linear AR model data, the data are noise free and \( f_\varepsilon = 0.99^4 \), the behavior of the AC of the original data and the surrogate data is very similar and the AC of the original data at time lag 1 falls within the distribution of the surrogate data like that shown in Figs. 7(d) and 7(e). The AMI of the original data falls within the distribution of the surrogate data like that shown in Fig. 7(f). Hence, we conclude that we cannot detect nonlinearity in irregular fluctuations: the correct result.

With irregular fluctuations from the Ikeda map data, noise free data and \( f_\varepsilon = 0.01 \), the AC of the original data is almost identical to that of the surrogate data and the AC of the original data at time lag 1 falls within the distribution of the surrogate data. The AMI of the original data falls outside the distribution of the surrogate data (like that shown in Fig. 7(c)). Hence, we consider that the irregular fluctuations include nonlinearity, which is the correct result. When data are contaminated by 10% observational noise, we can still detect linearity and nonlinearity in irregular fluctuations correctly by applying our method.

8. Applications

We now apply the proposed method to three experimental systems: (1) NMR laser data; (2) MGAT from September 1920 to December 2005 and (3) MSN from January 1749 to 10 August 2004. See Figs. 1(a-c), respectively. We use 2048 data points for the NMR laser data and the MSN data, and 1024 data points for the MGAT data. We apply the STFTS method to the NMR laser data because the data has no trend and the STFTS method to the MGAT and MSN data because the data have long term trends.

Figure 3 shows a segment of the surrogate data of the NMR laser data and the result, where \( f_\varepsilon = 0.1 \). From Fig. 3(a) we can know that there are some differences between the original and surrogate data, and Fig. 3(b) shows that the AMI of the NMR laser data falls outside the distributions of the surrogate data. Hence, we consider that the NMR laser data are nonlinear. This result is in agreement with the previously obtained understanding of the data [6].

We apply the STFTS method to the MGAT and MSN data. We increment \( f_\varepsilon \) in steps of 0.05. To know a rough upper bound of \( f_\varepsilon \) we first estimate the power spectrum of the data. Figure 4 shows that the power seems to be white when the frequency is larger than around 100 for both data sets. The result indicates that the rough upper bound of \( f_\varepsilon \) is 0.8 for the MGAT data and 0.9 for the MSN data.

We find that the AC of the MGAT and the MSN data at time lag 1 falls outside the distribution of the surrogate data when \( f_\varepsilon = 0.6 \) and \( f_\varepsilon = 0.3 \), respectively. These values are smaller than both the upper bounds. We show all results in
Table 1: Frequency domain when the NH is rejected and not rejected. The R indicates that the NH is rejected and the NR indicates not rejected. Hence, R implies that our method detects nonlinearity in the irregular fluctuations, and NR implies that our method fails to detect it. In all results shown in this table, the AC of the original data at time lag 1 falls within the distribution of surrogate data.

<table>
<thead>
<tr>
<th></th>
<th>Frequency domain $f_e$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NR</td>
<td>R</td>
</tr>
<tr>
<td>MGAT data</td>
<td>0.05 ~ 0.3</td>
<td>0.35 ~ 0.55</td>
</tr>
<tr>
<td>MSN data</td>
<td>0.05 ~ 0.25</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Surrogate data of time series shown in Fig. 1(b) and 1(c). (a) MGAT data and (b) MSN data.

Figure 6: An enlargement of the original data and one of the surrogate data. (a) MGAT data and (b) MSN data, where the solid line is the original data and the dotted line is the surrogate data.

Table 1. As mentioned previously, we adopt the last result. Hence, we show results for the MGAT data when $f_e = 0.55$ and for the MSN data when $f_e = 0.25$. Figure 5(a) and 5(b) show very similar behavior to Fig. 1(b) and 1(c), and Fig. 6 shows that local structures are different between the two. Figure 7(a) and 7(d) show that the AC of the original data falls within the distribution of the surrogate data in both cases. Figure 7(b) and 7(e) show that the AC of the original data is almost identical to the surrogate data. From these figures we conclude that linearity and long term trends are preserved in the surrogate data. From Figs. 7(c) and 7(f) we consider that we can detect nonlinearity in irregular fluctuations of the MGAT data and we cannot detect that of the MSN data.

9. Conclusion

We describe a method for investigating nonlinearity in irregular fluctuations, even if they exhibit long term trends. Our arguments and computational examples show that this algorithm succeeds in testing nonlinearity and discriminating well between linear and nonlinear data.

This research is supported by a Hong Kong University Grants Council Competitive Earmarked Research Grant (CERG) No. PolyU 5216/04E.

References

Inferring on the Dynamical Nature of Spike Trains

Oscar De Feo†,‡
†Department of Microelectronic Engineering, University College Cork
Butler Building, North Mall, Cork, Ireland
‡Laboratory of Nonlinear Systems, Swiss Federal Institute of Technology Lausanne
EPFL IC LANOS, Station 14, 1015 Lausanne, Switzerland
Email: oscar.defeo@ucc.ie

Abstract—By means of an ad hoc technique, based on the gray-box identification of an externally forced neuron model, we address the inference of the intrinsic deterministic dynamical nature of recorded in vivo neural activity, a far to be solved recurrent problem in experimental and theoretical neuroscience.

1. Introduction
With the availability of multielectrode measurement systems, which can measure the activity of hundreds of neurons at once, addressing the intrinsic deterministic dynamical nature of in vivo recorded neural activity—i.e. understanding if the intrinsic behaviour of a neuron would be quiescent (equilibrium), periodic (limit cycle), or irregular (chaotic), once that all the known possible external factors have been marginalised—has become a recurrent issue in experimental and theoretical neuroscience.

Although very important, the problem is very difficult and remains practically unsolved because, usually, the common experimental setups provides only spike trains, i.e. the time instants of spike occurrences (point process), without providing any direct insights about the sub-threshold and/or the membrane dynamics of the neurons.

Here, we transposed the problem to the inference of the qualitative behaviour of an externally forced gray-box neuron model fitted to the data, i.e. we fit a well-known neuron model to the the recorded data and, afterwards, we infer the qualitative behaviour of the free (unforced) model.

The method is explained in the next section then, after validation (cf. Sec. 3), is applied on real signals (cf. Sec. 4), providing new insights, which are discussed in Sec. 5.

2. Fitting Hindmarsh-Rose Neurons to Spike Trains

The problem can be formulated as follow:

H: given a set of input spike trains (external factors) and the response spike train of the neuron at issue,

T: determine the best (in predictive sense) parameters of a reference externally driven neuron model.

Namely, given multi-recorded spike trains, we assume one of them as the response and all the other as external factors to this one, including the external stimuli, to which we can associate a stimulation train, and afterwards we fit an a priori decided dynamical input-output relationship between them.

Given the point processes alone, the problem does not admit solution (cf. [6]). We circumvent that by means of an artificial reconstruction of the membrane potentials out of the measured spike trains, somehow similarly to [4].

We assume the Hindmarsh-Rose neuron [3] as reference model; though, we can proceed similarly for any other deterministic parametric model; then, the whole identification method is composed of three steps:

1: categorisation of the spikes within inputs and response spike trains;
2: artificial composition of the membrane potentials associated to inputs and response spike train;
3: identification of the parameters of an externally driven Hindmarsh-Rose neuron model out of the artificial inputs and response membrane potentials.

These steps are described in more detail in the following sections.

2.1. Step 1: Spikes Categorisation

The spikes of a given spike train are categorised into four possible classes according to the antecedent and consequent Inter-Spike Intervals (ISI), as shown in Fig. 1.

The precise values separating ISI in small and large varies from case to case, and has to be decided according to the measurements, for instance on the base of the ISI histograms so to separate the modes.

2.2. Step 2: Membrane Potentials Composition

Once that a spike train is classified according to the above algorithm, we compose an artificial membrane potential time series as follow. First, according to the classification, we paste one, out of four possible a priori determined spike shapes, at every spike time stamp. Then, the whole membrane potential is interpolated over all time instants by means of a constrained spline interpolation, i.e. under the constraint of only one minima between every two spikes.
The spike shapes used (from the Hindmarsh-Rose neuron model) are shown in Fig. 2, whilst a graphic illustration of the membrane potential construction is given in Fig. 3.

2.3. Step 3: Gray-box Model Identification

Finally, a driven Hindmarsh-Rose neuron model is fitted to the data obtained through the previous two steps. Namely, called \( y(t) \) the artificial membrane potential time course of the neuron at issue (the one that we want to identify) and \( u(t) \) the artificial membrane potentials of all the other concurrently measured neurons, including the stimulus which is considered as a fake neuron, we fit them to the reference model illustrated in Fig. 4.

In other words, the data \( \{y(t), u_k(t)\} \) are used to fit the parameters of the following non-autonomous ordinary differential equation (normalised electrically coupled Hindmarsh-Rose neuron):

\[
\begin{align*}
\dot{x}_1 &= ax_1^2 - x_1^3 - x_2 - x_3 + \sum_{k=1}^{K} A_k (u_k - y) \\
\dot{x}_2 &= (a + \alpha) x_1^2 - x_2 \\
\dot{x}_3 &= \mu (b x_1 + c - x_3) \\
y(t) &= x_1(t)
\end{align*}
\]

where the known data are the \( u_k(t) \) and \( x_1(t) \) (which is equal to \( y(t) \)), and the unknown (regressed) variables are \( a, b, c, \alpha, \mu, A_k, x_2(t) \) and \( x_3(t) \).

Despite intricate, the identification of model Eq. (1) is not very complex, it is just computationally expensive and tricky because of the slow-fast nature of the model [1].

Indeed, the Jacobian of system (1) is given by

\[
\begin{pmatrix}
(2a - 3x_1) x_1 - \sum_{k=1}^{K} A_k & -1 & -1 \\
2(a + \alpha) x_1 & -1 & 0 \\
\mu b & 0 & -\mu c
\end{pmatrix}
\]

and depends only on the \( x_1 \) state variable, which is known. Therefore, knowing \( x_1 \), system (1) can be rewritten as

![Figure 2: Spike shapes in normalised units of attack (a, blue), burst (b, red), decay (c, green), and lonely (d, magenta) spikes.](image)

![Figure 3: Artificial membrane potential construction: (a) – classified spikes, colours as above; (b) – \textit{a priori} spike shapes pasted at the corresponding time stamps; (c) – interpolated artificial membrane potential.](image)

![Figure 4: Reference model for the gray-box Hindmarsh-Rose neuron model.](image)

(a) (b) (c) (d)

Figure 1: Spikes categorisation according to the neighbouring ISI: (a) – classification table; (b) – example.

3. Method Assessment on Artificial Data

The aim of the method is assessing the qualitative \textit{intrinsic} nature of measured spike trains marginalising the further available measurements. Hence, we evaluated the method in this sense considering three significant cases (Fig. 5), \textit{i.e.} either a quite, a periodic, or a chaotic neuron driven by a \textit{measured} regular spike train, mimicking a stimulus-like input, and an \textit{unmeasured} Poissonian spike...
neuron, mimicking the background noise and the unmeasured inputs. Segments of the responses are shown in Fig. 6.

We applied the method assuming only the measured spike trains to be known with successful results. Indeed, as shown Fig. 7, the free identified neurons had intrinsic qualitative behaviour compatible with the measured ones. However, note that the parameters identified have no, nor they pretend to have, relation with those of the original (identifying) systems. Indeed, because of the artificial assembly of the membrane potentials, the identified and identifying parameters may be completely uncorrelated. Though, the intrinsic behaviour of the identified systems has shown to be correctly related to the one of the original system.

4. Application on Extra Cellular Spike Recordings

We applied the method on real data, i.e. in vivo multiunit spike train recordings from the brainstem trigeminal nuclei\(^1\), with the aim of characterising the intrinsic qualitative behaviour of sustainedly spiking neurons (about 50). For each identification, we assumed a sustained spike train as the response, and all the other spike trains concurrently measured, including the external stimulus, as inputs. An example of the considered data is reported in Fig. 8.

The summary of the results of the categorisation of the intrinsic qualitative asymptotic behaviour of the identified neurons is reported in Tab. 1. These results have been verified to be statistically significant in the way explained in the next section.

<table>
<thead>
<tr>
<th>steady</th>
<th>periodic</th>
<th>chaotic</th>
<th>undetermined</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.25%</td>
<td>36.17%</td>
<td>44.68%</td>
<td>14.90%</td>
</tr>
</tbody>
</table>

Table 1: Categorisation results for the intrinsic qualitative asymptotic behaviour of sustainedly spiking neurons.

4.1. Statistical validation

The categorisation of the asymptotic behaviour reported in Tab. 1 has been validated to be statistically significant as follow (random truncation statistic [2]). For each of the neurons at issue 20 independent segments of different lengths and positions have been drawn at random. Afterwards, the identification method has been applied to each segment, and the qualitative asymptotic behaviour of the identified system determined. Finally, the categorisation of the whole spike train has been validated with statistical significance \( p \geq 95\% \) verifying that at least 19 over 20 segments were classified as the whole segment. Whenever the statistical significance has not been verified, the asymptotic behaviour has been classified as undetermined.

\(^1\)Experiments and data descriptions available at http://aperest.epfl.ch.
5. Concluding Remarks

We proposed a method for addressing the intrinsic dynamical nature of in vivo recorded neural activity based on the gray-box identification of a neuron model.

After successful validation, the analysis of real data has shown that a large part, though not the entirety, of the sustainedly spiking neurons may be considered of chaotic nature.

The results give experimental evidence for chaotically behaving neurons modulated/controlled by external stimuli, which supports an alternative perspective of neural information processing, where a combination of chaotic behaviour and neural synchronisation provide at once feature binding together with the capacity of representing the diversity of external stimuli, as recently proposed by the APEREST project2.

Acknowledgments

O. De Feo acknowledge financial support from the EC-Project APEREST (IST-2001-34893 and OFES-01.0456), and is grateful to Martin Hasler and Valeri Makarov for the exciting and constructive discussions.

References


2http://aperest.epfl.ch.
Phase space projection with time domain constraint (TDC) for noise reduction

Xiaodong Luo, Jie Zhang and Michael Small

Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hung Hom, Hong Kong.
Email: enxdluo@eie.polyu.edu.hk, enzhangjie@eie.polyu.edu.hk, ensmall@polyu.edu.hk

Abstract—Phase space projection is a popular technique for speech argument. For practical purpose, it was proposed to impose a time domain constraint on the residual noise to obtain admissible communication quality. In this work we re-examine this technique. A more general version is derived without assuming the independence between the clean signal and the noise but includes the available techniques as subcases. We demonstrate the performance of this technique for noise reduction through the simulation data from the Rössler system and experimental speech record. Numerical results show that our algorithm has succeeded in augmenting the signal-to-noise ratio (SNR).

1. Introduction

Phase space projection is a popular technique adopted for noise reduction in various fields because of its simplicity and efficiency. For example, to reduce the component of white noise in a time series, it is suggested to extract qualitative dynamics from experimental noisy time series by removing the empirical orthogonal functions (EOFs) that correspond to the projections of the noise component in the trajectory matrix [2, 14]. However, it will become much more complicated to apply this technique if the data set is contaminated with colored noise. Distinction between the projections of the signal component and the noise component in the original coordinate space usually obscures. As a remedy, it is recommended to pre-whiten the colored noise by introducing a transformation to the covariance matrix of the noise component [1].

It is worth to note that, in the above two scenarios, noise reduction based on the phase space projection scheme will not operate on the EOFs that span the signal-plus-noise subspace, therefore those operations could achieve a lower possible distortion for the clean signal, but at the price of a higher possible residual noise level [4]. In the community of speech communication, the residual noise is interpreted as the background noise, while the clean signal is the message to be transferred. Of course, too high residual noise is not desirable since it will interfere the communication between the speakers. A solution to this problem is to allow certain distortion of the clean signal so that the residual noise is reduced to an admissible level. As proposed in the literature (e.g., [4]), a time domain constraint (TDC) is introduced to impose on the residual noise. And then the problem is to find the optimal tradeoff between signal distortion and residual noise so as to minimize the overall distortion (as a measure of the communication quality).

In the literature, both the situations of a data set contaminated with white and colored noise are studied, see, for example, [3, 4, 7]. However, in those works, usually the authors make two assumptions about the experimental time series. The first assumption is that the time series is stationary and ergodic, and the second is that the noise components are independent of the clean signal. In this work we will re-examine the idea of the TDC projection and extend it to a more universal version. We will also show that, with the first assumption, the second is not necessary in general.

We will introduce the primary idea of TDC projection in section II. Given a trajectory matrix of a time series, its optimal TDC projector is obtained in the sense of minimizing the signal distortion subject to an admissible noise level. In section III we will apply the optimal TDC projection technique to the simulated data of the Rössler system and an experimental speech data. We will also compare the performance of the projections under different TDCs. Finally, a conclusion is available to close the whole work.

2. Optimal TDC projection

Suppose that a noisy time series \( s = \{s_i\}_{i=1}^M \) consists of the clean signal component \( d = \{d_i\}_{i=1}^M \) and the additive noise component \( n = \{n_i\}_{i=1}^M \), and for each data point \( s_i \), one has \( s_i = d_i + n_i \). In addition, following the assumption in the literature it is assumed that the data set \( \{s_i\}_{i=1}^M \) is (weakly) stationary and ergodic so that its expectation exists and its variance is finite, while its (auto)covariances only depend on the time difference between the subsets.

Based on the time series \( \{s_i\}_{i=1}^M \), one can construct a \((M - m + 1) \times m\) matrix

\[
S = \begin{pmatrix}
S_1 & S_2 & \ldots & S_m \\
S_2 & S_3 & \ldots & S_{m+1} \\
\vdots & \vdots & \ddots & \vdots \\
S_{M-m+1} & S_{M-m+2} & \ldots & S_M
\end{pmatrix}_{(M-m+1) \times m}, \tag{1}
\]

which will be called the trajectory matrix of the time series \( s \) hereafter. In a similar way, one can also construct the tra-
jectory matrices $D$ and $N$ corresponding to the components $\{d_i\}_M$ and $\{n_i\}_M$ respectively. And it is easy to verify that $S = D + N$.

Because of the presence of noise, usually $S$ is a full rank matrix. Without loss of generality, let us suppose $m < (M - m + 1)$. Then the operation of phase space projection occurs in the $m$-dimensional space. By introducing a projection operator $H$ to the trajectory matrix $S$, one obtains the matrix $Z = SH$. Let $R_m = Z - D = D(H - I_m) + NH$ be the matrix of residual signal, where the term $D(H - I_m)$ corresponds to signal distortion and the term $NH$ to residual noise. For the purpose of noise reduction, it is required to achieve as small signal distortion as possible. Thus letting the operator $H = I_m$ (the identity matrix) is an intuitive choice. However, for speech communication it is also demanded that the residual noise reduces to an admissible level. Therefore the objective now shall be to minimize signal distortion subject to the achievement of a permissible residual noise level. For this reason, the choice of $H = I_m$ might not fulfill this requirement, one may actually further reduce the residual noise level at the price of introducing certain signal distortion, which is the focus of this study.

Similar to the idea in the work [4], let us impose a time domain constraint (TDC) $\mu$ on the term of residual noise $NH$ and treat $R = D(H - I_m) + \mu NH$ as the part that requires an overall minimal distortion, where $\mu^2 \in [0, +\infty)$ is the Lagrange multiplier determined by the admissible noise level from the practical demand (see the discussions in [4]). As the measure of the overall distortion, the average energy

$$
\Xi \approx \frac{1}{(M - m + 1)m} tr(R^T R) \quad (2)
$$

is the target for minimization, where $tr(\cdot)$ means the trace of a square matrix, and $R^T$ denotes the transpose of the matrix $R$ [10].

With $R = D(H - I_m) + \mu NH$, it can be shown that

$$
tr(R^T R) = tr(H^T (D + \mu N)^T (D + \mu N) H) - 2 tr((H^T (D + \mu N)^T D + D^T (D + \mu N)^T) \quad (3)
$$

Thus the minimization of the average energy (Eq. (2)) requires that $\partial tr(R^T R)/\partial H = 0$, whose solution [11] leads to the optimal projector

$$
H_{min} = (D + \mu N)^T (D + \mu N)^{-1} \quad (D + \mu N)^T D \quad (4)
$$

With the noise components, $\partial tr(R^T R)/\partial H^T = 2(D + \mu N)^T (D + \mu N)$ is positive definite, which confirms that the extremum taken at $H_{min}$ is a minimum.

Since $S = D + N$, one can also express Eq. (4) in terms of $S$ and $N$, i.e.,

$$
H_{min} = (S + (\mu - 1) N)^T (S + (\mu - 1) N)^{-1} \times (S + (\mu - 1) N)^T (S - N) \quad (5)
$$

As one may notice, the derived result only relies on the assumption that the time series is stationary and ergodic, and the condition of independence between the clean signal $d$ and the noise component $n$ is not necessary. If under the assumption that the clean signal $d$ and the noise component $n$ are independent, statistically one has $D = N^T = 0$ as $M \to \infty$. Therefore, $S^T S = D^T D + N^T N$, and Eq. (5) reduces to

$$
H_{min} = (S^T S + (\mu^2 - 1) N^T N)^{-1} (S^T S - N^T N), \quad (6)
$$

which is the main result in the literature, for example, [3, 4, 7, 8]. In this sense, it is clear that Eq. (5) is a more general version than Eq. (6).

3. Applications

It is worth to note that the trajectory matrices are all Hankel matrices. For example, consider the matrix $S$ in Eq. (1), its entries satisfy that $S(i, j) = S(k, l)$ if $i + j = k + l$, where $S(i, j)$ is the element of matrix $S$ on $i$-th row and $j$-th column. However, the transformed matrix $Z = SH$ usually is not a Hankel matrix, and there are many choices to extract the augmented (scalar) data $[z_i^M]_{i=1}$ from the matrix $Z$. In this work we adopt the method of secondary diagonal averaging [5, p. 24], which takes the average over the elements along the secondary diagonals of matrix $Z$ as the filtered signal $[z_i]_{i=1}^M$, and thus can form a new trajectory (Hankel) matrix $Z^H$ from $[z_i^M]_{i=1}$. The main advantage of this choice is that it is optimal among all “Hankelization” procedures in the sense that the matrix difference $Z^H - Z$ has minimal Frobenius norm [5, p. 24 & p. 266].

We select the signal-to-noise ratio (SNR) as the measure to evaluate the performance of our scheme for noise reduction, which is defined (in dB) as [4, 8]

$$
SNR = 10 \log_{10} \frac{||d||^2}{||z - d||^2}, \quad (7)
$$

where $||d||^2 = \sum_{i=1}^M d_i^2$ and $||z - d||^2 = \sum_{i=1}^M (z_i - d_i)^2$. From the definition, it can be shown that $||d||^2 / ||z - d||^2$ is the ratio of the average energy of the clean signal to that of the residual signal (the overall distortion), and minimizing the average energy of the residual signal will optimize the SNR performance.

We first apply our method to a simulated time series (the $x$ component) of the Rössler system

$$
\begin{align*}
\dot{x} & = -(y + z) \\
\dot{y} & = x + 0.15y \\
\dot{z} & = 0.2 + (x - 10)z
\end{align*} \quad (8)
$$

The sampling time is 0.1 units for the integration of the differential equations. A time series with 9,000 data points is generated after sufficiently long time to avoid transient states. To construct the trajectory matrices, the window size is set to $m = 20$. 

360
Table 1: Performances of the TDC projections for the Rössler system in terms of SNR (dB) (reproduced from [10]).

<table>
<thead>
<tr>
<th>TDC µ</th>
<th>Additive white noise</th>
<th>Additive colored noise</th>
<th>Multiplicative noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>20 → 25.50 ± 0.09</td>
<td>20 → 20.92 ± 0.05</td>
<td>20 → 21.11 ± 0.10</td>
</tr>
<tr>
<td></td>
<td>10 → 15.95 ± 0.11</td>
<td>10 → 10.89 ± 0.04</td>
<td>10 → 11.14 ± 0.09</td>
</tr>
<tr>
<td></td>
<td>0 → 5.88 ± 0.06</td>
<td>0 → 0.87 ± 0.04</td>
<td>0 → 1.16 ± 0.10</td>
</tr>
<tr>
<td>0.5</td>
<td>20 → 25.80 ± 0.10</td>
<td>20 → 21.07 ± 0.05</td>
<td>20 → 23.23 ± 0.24</td>
</tr>
<tr>
<td></td>
<td>10 → 17.74 ± 0.15</td>
<td>10 → 11.64 ± 0.06</td>
<td>10 → 14.42 ± 0.23</td>
</tr>
<tr>
<td></td>
<td>0 → 9.71 ± 0.10</td>
<td>0 → 3.12 ± 0.08</td>
<td>0 → 6.97 ± 0.22</td>
</tr>
<tr>
<td>1.0</td>
<td>20 → 26.27 ± 0.11</td>
<td>20 → 21.17 ± 0.05</td>
<td>20 → 24.44 ± 0.34</td>
</tr>
<tr>
<td></td>
<td>10 → 18.29 ± 0.16</td>
<td>10 → 11.89 ± 0.07</td>
<td>10 → 16.12 ± 0.32</td>
</tr>
<tr>
<td></td>
<td>0 → 10.10 ± 0.09</td>
<td>0 → 4.15 ± 0.09</td>
<td>0 → 9.56 ± 0.33</td>
</tr>
</tbody>
</table>

Now let \( \{s_i\}_{i=1}^M \) and \( \{d_i\}_{i=1}^M \) denote the noisy and clean signals of the Rössler time series respectively. In the following let us consider three scenarios where the clear signal is contaminated with certain type of noise. In the first case, the contamination source is additive white noise \( \{\xi_i\}_{i=1}^M \) (so that \( s_i = d_i + \xi_i \)), which follows the normal Gaussian distribution \( N(0, 1) \) (denoted by \( \xi_i \sim N(0, 1) \) henceforth). In the second case, the contamination is additive colored noise \( \{\eta_i\}_{i=1}^M \) (so that \( s_i = d_i + \eta_i \)), which is produced from the third order autoregressive (AR(3)) process \( \eta_i = 0.8\eta_{i-1} - 0.5\eta_{i-2} + 0.6\eta_{i-3} + \xi_i \) (\( \xi_i \sim N(0, 1) \)). In the last case, the contamination source is multiplicative noise \( \{\zeta_i d_i\}_{i=1}^M \) (so that \( s_i = (1 + \zeta_i)d_i \)). As an example, we let \( \zeta_i = \eta_{i}^2 \), where \( \{\eta_i\}_{i=1}^M \) is from the previous AR(3) process. Obviously, in this case the noise component \( \{\zeta_i d_i\}_{i=1}^M \) is correlated to the clean data \( \{d_i\}_{i=1}^M \).

For each of the above cases, the contamination noise has three different levels (20 dB, 10 dB, 0 dB respectively), and for each noise level, we generate 10 different noise samples from the same process for calculation. Our objective is to study the performance of phase space projection for noise reduction under different TDC constraints \( \mu \). As examples, we let TDC \( \mu = 0, 0.5 \) and 1 separately. TDC \( \mu = 0 \) corresponds to the least-squares (LS) projection based on the SVD technique that has been studied in, for example, [1, 2, 14] \(^2\). While \( \mu = 1 \) leads to the well-know linear minimum mean-squared-error (LMMSE) projection [8, 13]. For comparison, the performances of these TDC projections are listed in Table 1 in terms of SNR values.

From Table 1 it is easy to find that, for the Rössler system our method works for all of the three scenarios. But the effect of data augmentation for additive colored noise is not as obvious as those for additive white noise and multiplicative noise. It also shows that, in general, the LMMSE projection has superior performance to that of the LS projection in the measure of SNR.

We then apply our method to an experimental speech record with 8,000 data points, which is sampled at 44 kHz and quantized to 16 bits (Fig. 1 (a)). But note that, in this case for the noisy signal, the noise is not separable from the clean signal (without the \textit{a priori} knowledge), therefore it would be preferred if one could produce a set of realizations that mimic the behavior of the underlying noise. For this purpose, the pseudo-periodic surrogate (PPS) algorithm [12] is adopted to generate 9 surrogates based on the original background noise (in the period without the speech signal). Our calculations reveal that, the SNR of the original speech data is about \( -0.32 \pm 0.18 \) dB. After introducing phase space projections to the speech data, for the LS (\( \mu = 0 \)) projection \(^3\), the augmented SNR\( \approx 4.36 \pm 0.41 \) dB. While for the cases of TDC \( \mu = 0.5 \) and 1, the corresponding SNRs increase to \( 6.28 \pm 0.61 \) dB and \( 6.97 \pm 0.66 \) dB respectively. For illustration, we plot the waveforms of the original speech record and the three augmented data under different TDCs in Fig. (1). The similar result is obtained in this case, that is, the LMMSE projection (\( \mu = 1 \)) produces a smoother speech waveform (panel (d)) than that of the LS projection (panel (b)), which is an advantage against its rival in speech communication since it often means better communication quality.

4. Conclusion

In this work we re-examined the phase space projection technique for noise reduction, especially in the field of speech communication. By imposing a time domain constraint on the residual noise, the optimal projection is obtained in the sense of minimizing the overall signal distortion (subject to an admissible noise level). It was also showed that, in general one need not assume independence between the clean signal and the noise component as did in

\(^2\)In this case one has to specify the dimension of the signal-plus-noise subspace in order to group the EOFs and eigenvalues that correspond to the noisy signal and remove the complementary noise subspace. Here we adopt the criterion of false nearest neighbor [9] implemented in the TISEAN package [6] and set the dimension size \( K \) to be 5.

\(^3\)The window size \( m = 30 \) and the dimension size of signal-plus-noise subspace to be \( K = 8 \)
Figure 1: (a) Original speech record; (b), (c) and (d) Speech data output from TDC projectors with $\mu = 0, 0.5$ and 1 separately (reproduced from [10]).

the literature. This viewpoint was confirmed by our numerical results.

Acknowledgments

This work was supported by a direct allocation grant from Hong Kong polytechnic University (A-PF95).

References

Predictive models of wind for controlling a wind turbine

Yoshito Hirata1, Hideyuki Suzuki1,2, and Kazuyuki Aihara1,2,3

1 Department of Mathematical Informatics, The University of Tokyo, Tokyo 113-8656, Japan
2 Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
3 Aihara Complexity Modelling Project, ERATO, JST, Japan
Email: yoshito@sat.t.u-tokyo.ac.jp

Abstract—We are considering generating more electricity by adjusting wind turbines to the future wind. By formulating this problem mathematically, we show that for controlling wind turbines, we need to predict the wind direction and the absolute wind speed for certain steps. For this problem of multi-step predictions, we compare several possible methods including direct and iterative predictions.

1. Predicting the wind

Since the number of wind turbines is increasing worldwide, predicting the wind is getting much attention. There are mainly three purposes for predicting the wind. The first purpose is to estimate the future supply for the electricity. For this purpose, we need to predict the wind in the order of hours and it is done by medium-range weather forecasts [1] or time series predictions [2, 3, 4]. The second purpose is to avoid damages caused by gusts. For this purpose, we need to predict the timing of gusts in the second order. It is done by a Markov chain [5, 6, 7, 8].

The third purpose is to generate more electricity by adjusting a wind turbine to the predicted wind. For this purpose, we need to predict the wind direction and absolute wind speed for a certain period of time. We formulate this problem of controlling a wind turbine as an optimal control that maximises the net benefits generated by a wind turbine. In detail, the problem can be described as follows: Let \( \hat{v}_t \) be the expected absolute wind speed at time \( t \), \( \hat{\theta}_t \), the expected wind direction at time \( t \), \( \phi(t) \), the controlled normal direction of wind turbine, \( f(w) \), the benefit obtained when the wind speed perpendicular to the face of the rotor is \( w \), and \( c(\phi) \), the total costs necessary for controlling the wind turbine along the orbit \( \phi \). Then the problem to solve is written as

\[
\max_{\phi} \left\{ \sum_{t=1}^{k+T-1} f(\hat{v}_t \cos(\hat{\theta}_t - \phi(t))) - c(\phi) \right\}.
\] (1)

The term \( c(\phi) \) can be further decomposed into, for example, three terms such as

\[
C_1 \sum_{t=k+1}^{k+T-1} ||\phi(t) - \phi(t-1)|| + C_2 \sum_{t=k+1}^{k+T-1} ||\phi(t) - \phi(t-1)||
+ C_3 \sum_{t=k+1}^{k+T-1} ||\phi(t) - \hat{\theta}_t||.
\]

Figure 1: Observational setups for (a) 23 April 2006 and (b) 29 April 2006.

The first term is proportional to how big angle we turn the wind turbine by. The second term is for changing the turning velocity. The third term is for damages caused by the difference between the wind direction and the direction of wind turbine.

2. Experimental observation

For establishing some methods for predicting the wind direction and absolute wind speed, we used real wind data. The datasets we use here were observed with 50 Hz for about 6 hours in the inner court of our institute on 23 April 2006 and 29 April 2006. We located two identical anemometers about 1 m above from the ground. On 23 April 2006, an anemometer was located 5 m west from the other, while it was located 5 m south from the other on 29 April 2006 (Fig. 1). The dominant wind direction on the both days was that of west. We consider these settings as a 1/100 model of a wind farm.

In the previous studies, we found that the wind measured in a similar way has serial dependence and that it is sometimes nonlinear [9]. During the test for serial dependence, a rejection occurred in a strange way. Our analyses [10] using models confirmed that this might happen by observational noise, dynamical noise, and/or high dimensionality of system. We also tried to predict the wind speed for a certain direction and showed that the wind speed is not predicted well using the observation of the predicted point only, while it can be better predicted if we have an extra observation point at the upstream [9].
Since the datasets look contaminated with observation noise, we took the moving average of 2 seconds and re-sampled it every 2 seconds. In what follows, we try several methods for predicting the wind direction or the absolute wind speed at an anemometer. During the evaluation, we used the first 2000 points for making a mathematical model and tried to predict the following 200 points. Then we shifted the windows by 200 points and repeated the above processes 44 times. We calculated the root mean square errors for the whole of each data set.

3. Direct and iterative methods

For predicting the wind direction or the absolute wind speed, we consider two classes of predictions, each of which contains two methods. The first class is that of direct predictions. The first method in this class is the speed and direction (sd) method. In this method, we prepare the wind directions and absolute wind speeds of the past and predict the future wind direction or absolute wind speed from them directly. But this method did not work well for the wind direction since the wind direction takes a value on a ring [11, 12]. Our second method in this class is the 2 dimensional (2d) method. In this method, we use the east and north winds of the past to predict the future values of these directly. Then we calculate the wind direction from them.

The second class is that of iterative predictions. The first method in this class is the simple iterative method. In the simple iterative method, we prepare a model predicting for a short range and apply it several times to have the predictions of the east and north winds at both points for appropriate steps. Then we calculated the quantities we want from them. The model for predicting for the shortest term is equivalent to that of 2d method. We cannot use the model for the sd method here because the sd method does not yield good predictions for the wind direction. There is a long debate on direct and iterative predictions [13, 14, 15]. One of the products is the second method in this class, the $\xi\phi$-method [15]. In this method, first we predict the future values using the simple iterative method. Then we correct them using another function called $\xi$. Lastly, we calculated the wind direction or absolute wind speed from them.

When building each model, first we selected an optimal set of delays for reconstruction using the method of Ref. [11] and then we constructed a radial basis function-affine model as prescribed as in Ref. [17].

4. Predicting the wind direction

We first predicted the wind direction using the above methods. The prediction errors are summarised in Fig. 2. The sd method was always the worst as we expected. The iterative methods worked well. Especially the $\xi\phi$-method achieved one of the smallest errors among the tested methods. When the second observation point was neither at the upstream nor at the downstream, the 2d method performed similarly as the persistent prediction, where we declared the value 2 seconds before as a prediction. In Ref. [15], it was argued that a direct prediction is better when the model predicting a short term is not accurate or there is a dynamical noise. Our results might imply that we have a relatively good short-term model and that the effects of dynamical noise is not so significant.

5. Predicting the absolute wind speed

We also predicted the absolute wind speed. The performances are compared in Fig. 3. The sd method worked better than the others especially when we predicted the absolute wind speed for long ranges. It also worked well when the second observation point was neither at the upstream nor at the downstream. These results simply seem to mean that in general we had better build a model that predicts what we want.

6. Conclusion

For predicting the wind direction, we can enjoy the better predictions by first predicting the wind speed in each direction and then calculating the interested quantity from these predictions. The iterative predictions are better than the direct predictions especially when the second observation point is at neither the upstream nor the downstream. For predicting the absolute wind speed, it is better to use the direct and speed method especially when the second observation point is at neither the upstream nor the downstream.

Thus we should combine the two above approaches for simulating controlling a wind turbine.

Acknowledgement

This study was partly supported by the Industrial Technology Research Grant Program in 2003, from the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

References


Figure 2: Comparisons of different prediction methods in root mean square errors for predicting the wind direction. Figures (a) and (b) were obtained from the dataset taken on 23 April 2006, and Figs. (c) and (d), from the dataset taken on 29 April 2006. In each figure, the horizontal axis corresponds to the prediction steps, i.e., how far future we predicted, and the vertical axis shows the root mean square error. The green, blue, red, magenta, and black lines indicate the performances of sd method, 2d method, simple iterative method, $\xi\phi$-method, and persistent prediction, respectively.


[10] Y. Hirata, S. Horai, H. Suzuki, and K. Aihara, “Wayland method can rate some datasets as less deterministic than their random shuffle surrogates,” Mathe-
Figure 3: We predicted the absolute wind speed using different prediction methods and compared their root mean square errors. Figures (a) and (b) were for the dataset taken on 23 April 2006, and Figs. (c) and (d), for the dataset taken on 29 April 2006. In each figure, the horizontal axis indicates how far future we predicted, and the vertical axis, the root mean square error. The green, blue, red, magenta, and black lines show the performances of sd method, 2d method, simple iterative method, $\xi\phi$-method, and persistent prediction, respectively.


Chaotic correlation among cycles in human electrocardiograms

J. Zhang, X. Luo and M. Small

Department of Electronic and Information Engineering,
Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong
Email: enzhangjie@eie.polyu.edu.hk, enxdluo@eie.polyu.edu.hk emsmall@polyu.edu.hk

Abstract—By using a recently developed method that is specifically designed to seek temporal correlation among cycles in pseudoperiodic time series, we are able to identify chaos in the human electrocardiogram (ECG) from both healthy and pathological groups. This method uses correlation coefficient as a measure of distance between cycles and further constructs a hierarchy of pseudo cycle series that in turn preserve less determinism than the original one. Appropriate statistics are then devised to reveal the temporal and spatial correlation encoded in this hierarchy of pseudo cycle series, which allows for reliable detection of chaos in the original time series. The method is free from phase space reconstruction and also proved to be robust to different noise and nonstationarity. We provide evidence that the human cardiac system is chaotic by using this method.

1. Introduction

During the last two decades, a great deal of work has been devoted to understanding the physiological information behind the variability of the cardiac cycle. Among them, nonlinear system theory has provided a new paradigm to explain the seemingly random signal and appears to be very promising in understanding the complicated dynamics in the human heart. However, the number of successful practical applications of chaotic methods to biological data has remained rather small. This is because a typical biological data set will be short, noisy and nonstationary enough to yield meaningful results. Classical dynamical indices therefrom such as correlation dimension, Lyapunov exponent and entropy [1, 2, 3] have been used extensively to examine whether the human electrocardiogram (ECG) time series is chaotic. However, the reliability of these indices has been questioned. The basic difficulty is in ascertaining whether the experimental time series is generated by a chaotic or a linear stochastic process. It has now been realized that the usual measures like saturation of correlation dimension and existence of positive Lyapunov exponent cannot by themselves establish the chaotic behavior of the system [4].

In contrast to the above nonlinear approaches, the surrogate data method [5] provides a linear statistical test in search of nonlinear determinism. For example, Small et.al.[6, 7] have produced a novel pseudoperiodic surrogate and concluded that human ECG during both sinus rhythm and VT is consistent with deterministic nonperiodic inner-cycle dynamics. Surrogate data analysis together with the short-term prediction analysis are also used to check whether the time series fromm ECG is consistent with the hypothesis of deterministic chaos [8]. It should be noted that for surrogate data analysis, we usually cannot detect the presence of chaos directly, but claim that, for example, the underlying process is consistent with a nonlinear one.

In this paper, we apply a newly developed method [9] to the human ECGs to detect the presence of chaos directly. Unlike most methods that require the reconstruction of the phase space, the method used here divide the pseudoperiodic time series into successive cycles that serve as the basic processing units. Linear correlation coefficient is then used as a measure of the distance between different cycles and the cycles are rearranged accordingly, building a hierarchy of pseudo cycle series that effectively encodes the inner-correlation in the original time series. Several new statistics are then devised to extract information within and across these newly created pseudo cycle series, which enables the detection of determinism and chaos. This approach is demonstrated to be robust to both measurement noise (white and colored noise) and dynamical noise, and this can be of special benefit in ECG analysis, in which the time series are always noisy and nonstationary.

2. Description of the method

We use the chaotic time series from the x component of the well know Rössler system given by:

\[
\begin{align*}
    x' &= -(y + z) \\
    y' &= x + 0.398y \\
    z' &= 2 + z(x - 4)
\end{align*}
\] (1)

for illustration of the method. Given the pseudoperiodic time series \(x_i\) of \(n\) observations, the first step is to segment the pseudoperiodic time series into \(m\) consecutive cycles according to the local minimum (or maximum), denoted as \(\{C_1, C_2, ..., C_m\}\). For each pair of cycles \(C_i\) and \(C_j\) \(i, j = 1, 2, ..., m, i \neq j\) with length \(l_i\) and \(l_j\), respectively, we then define the correlation coefficient as follows (without loss of generality, suppose \(l_i < l_j\))

\[
\rho_{ij} = \max_{l_i \leq l \leq l_j} \frac{\text{Cov}[C_i(1 : l), C_j(1 + l : l + l)]}{\sqrt{\text{Var}[C_i(1 : l)]} \sqrt{\text{Var}[C_j(1 + l : l + l)]}}
\] (2)
where $C_i(a : b)$ denotes the segment between the $a$th and the $b$th element in $C_i$. This definition means that if two cycles are not of the same length, we will shift the shorter cycle $C_i$ onto the longer one $C_j$ for $(l_j - l_i)$ steps, calculate one correlation coefficient between $C_i$ and the corresponding part of $C_j$ in each step, and pick out the largest one as the correlation coefficient between $C_i$ and $C_j$.

The correlation coefficient $p_{ij}$ measures the similarity of waveform between cycle $C_i$ and $C_j$. The larger the $p_{ij}$, the higher the level of similarity. Considering the continuity and smoothness of the vector fields of deterministic systems, two cycles with a larger $p_{ij}$ will also be close in the phase space.

Now we study the correlation coefficient between each pair of cycles in detail. For cycle $C_i$, $(i = 1, 2, \ldots, m)$, we first calculate its correlation coefficients with the remaining $(m - 1)$ cycles $C_j$ $(j \neq i)$. Then we sort these $p_{ij}$s in descending order, and the $(m - 1)$ cycles are also rearranged correspondingly. Denote the sorted cycle sequence as a column vector $R_i = [C_{S_1}, C_{S_2}, \ldots, C_{S_{m-1}}, \cdots, C_{S_m}]$, where $S_{i,j}$ is the index of the $j$th most similar cycle to $C_i$. Then, by picking out the $p$th $(1 \leq p \leq m - 1)$ element from each column $R_1, R_2, \ldots, R_m$ and linking them together in order, we can build a (row) sequence of $m$ cycles, denoted as $T_p = [C_{S_{1,p}}, C_{S_{2,p}}, \ldots, C_{S_{m,p}}]$. For consistency, the original cycle series is denoted as $T_0 = [C_1, C_2, \ldots, C_m]$.

Note that each cycle in $T_p$ is the $p$th “closest” to the corresponding cycle in $T_0$, therefore as $p$ gradually increases, $T_p$ will grow less and less similar to $T_0$, i.e., $T_1$ is the most similar and nearest cycle series to $T_0$, while $T_{m-1}$ is the most different and farthest one. We call these $m - 1$ cycle series pseudo cycle series as opposed to the “real” cycle series $T_0$. The construction of this hierarchy of pseudo cycle series $T_p (p = 1, 2, \ldots, m - 1)$, as we shall see later, provides a novel way to examine the deterministic and chaotic structure in the original time series. It should be noted that the $T_p$'s are not necessarily reorderings of $T_0$, since, for example, two cycles may have the same cycle as nearest neighbor which would result in a double entry in $T_1$.

Now we demonstrate how to extract useful information from the $T_p$'s. For clarity of notation, we use $S_p$ to represent the sequence of the cycle indexes in $T_p$, i.e., $S_p = [S_{1,p}, S_{2,p}, \ldots, S_{m,p}]$. Then we count the number of cycle-pairs in $T_p$ that satisfy the following condition

$$S_p(i + k) - S_p(i) = k, (1 \leq i \leq m - k; k \geq 1) \quad (3)$$

where $S_p(j)$ represents the $j$th element in $S_p$. Physically, this means that we not only see cycle $C_i$ evolve into cycle $C_{i+k}$ after $k$ cycles in $T_0$ (which is trivial), but also see $C_i$'s $p$th closest cycle $C_{S_{i,p}}$ evolve into $C_{i+k}$'s $p$th closest cycle $C_{S_{i+k,p}}$ after $k$ cycles in $T_p$. Intuitively, this indicates the two cycles $C_i$ and $C_{S_{i,p}}$ nearby in phase space are strongly correlated by sharing similar dynamical evolution, and the correlation time lasts for $k$ cycles. We use $N_{pk}$ to denote the number of cycle-pairs in $T_p$ that satisfy the condition (3). Note that for higher $p$ and higher $k$ the condition (3) can be fulfilled by chance. To avoid this, we use a more strict condition, i.e., for each cycle span $k$ only the pairs of cycles should be counted that fulfill condition (3) for all cycle span $\leq k$. Considering it’s hard for this condition to be strictly met in the presence of noise, we suggest using this condition for clean data, and condition (3) for noisy data.

For chaotic systems, the distance between two nearby cycles will increase exponentially over time due to the very nature of sensitivity to initial conditions. Therefore, the correlation between two cycles, which is reflected in $N_{pk}$, is also expected to drop exponentially with the cycle span $k$. The semilog plot $\ln(N_{pk}) \sim k$ thus appears to be a straight line (See Fig.1), whose slope is actually related to the largest Lyapunov exponent. The larger the $[\ln(N_{pk})/k]$, the higher the level of chaos. So we can use $[\ln(N_{pk})/k]$ as an indicator of chaos, which we call cycle divergence rate (CDR). Usually we use $p = 1$ for CDR, because $T_1$ is the most similar cycle series to $T_0$, and therefore maintains most of the determinism and chaos. When $p$ increases, $T_p$ will inherit less determinism and chaos from $T_0$, therefore $\ln(N_{pk}) \sim k$ will be subject to more statistical fluctuations, and $[\ln(N_{pk})/k]$ (a linear fit to $\ln(N_{pk}) \sim k$) will also drop because the level of chaos decreases.

It’s interesting to see that by summing the $N_{pk} \sim k$ curve for the first $\theta$ pseudo cycle series, i.e., $N_\theta = \sum_{p=1}^{\theta} N_{pk}$ $(\theta = 0.15m)$, we can find a power law relation between $N_\theta$ and $k$, i.e., a linear dependency between $\ln(N_\theta)$ and $\ln(k)$. This is similar to the combination of a number of stochastic processes whose autocorrelation functions decay exponentially with different time constants, and the resultant process assumes a power law relation in its autocorrelation function. We call $[\ln(N_{pk})/k]$ average cycle divergence rate (ACDR), see Fig.2 (a) (the line with symbol “ring”) for illustration. And a larger ACDR indicates higher level of chaos. Experimentally, ACDR is more robust to noise. See Fig.2 for the ACDR in the presence of different kinds of
3. Application to human electrocardiograms

In this section, we calculate the ACDR for time series from various ECGs. The reason why we choose ACDR rather than CDR is that the former is more robust to different kind of noise (which is common for the ECG time series) since it involves a summation over all pseudo time series produced.

Fig.3 give the results for the ECGs being checked. We find that the ECGs from different subject, either healthy or those suffer from heart diseases, or even those in VF (ventricular fibrillation) all exhibit chaotic correlation between cycles, i.e., there are scaling regions in all the ACDR plot on a log-log scale. Actually for most ECG time series we have examined, the chaotic correlation between cycles is found to be present. This finding is consistent with previous results in the literature. Different from them, however, is that with our method chaos is detected directly by finding the correlation between cycles decreasing exponentially with the cycle span, rather than through surrogate data methods or using classical measures that easily lead to false results in the presence of noise. The finding of chaos in ECG has great clinical implication. It open the possibility to control the heart to desired state via the “chaos control”.

It should be noted that there are not significant difference in the slope of the scaling region in the ACDR plots across different groups. As for the ECG from ventricular fibrillation (VF), the corresponding time series is relatively short, thus leading to a short scaling range in ACDR.

4. Conclusion

By using a recently developed method that seeks chaotic correlation among cycles in pseudoperiodic time series, we have directly identified chaos in the human electrocardiograms for various groups including healthy, pathological, and those in VF. Our work suggest that the dynamics of heart might be amenable to the chaos control techniques.

Acknowledgments

This research was funded by a Hong Kong University Grants Council Grant Competitive Earmarked Research Grant (CERG) number PolyU 5235/03E.

References


Figure 2: ACDR for Rössler system with (a) additive Gaussian noise of different levels. (b) additive colored noise from the AR process $r_{n+1} = ar_n + b \eta_n$ ($\eta_n$ is the Gaussian noise term) when $a = 0.1, b = 0.3$ (noise level 4%), $a = 0.5, b = 0.8$ (noise level 12%), and $a = 0.8, b = 0.9$ (noise level 20%). (c) dynamical noise of different levels. Noise term $E \eta(t)$ is added to the right-hand side of the first equation in (1), which is integrated at a time span of 0.2.
Figure 3: ACDR for ECGs from (a) healthy volunteer, age 20, male (b) coronary care patient in normal state, age 59, female (c) coronary care patient in normal state, male (d) coronary care patient in VF, male.


The recurrence quantification technique applied to Liu’s system

Francisco Balibrea Gallego†, M. Victoria Caballero Pintado‡ and Lourdes Molera Peris‡

†Departamento de Matemáticas, Campus de Espinardo, 30100 Murcia, Spain
‡Departamento Métodos Cuantitativos para la Economía, Campus de Espinardo, 30100 Murcia, Spain
Email: balibrea@um.es, mvictori@um.es, lmolera@um.es

Abstract

Recurrence Quantification Analysis is used to detect transitions chaos to periodical states or chaos to chaos in a new dynamical system proposed by Liu et al. This system contains a control parameter in the second equation and was originally introduced to investigate the forming mechanism of the compound structure of the chaotic attractor which exists when the control parameter is zero.

1. Introduction

One common procedure, indeed the only one at times, to understand the behavior of a dynamical phenomenon is through the study of the time series it generates. This is particularly interesting in nonlinear cases, as is shown by its use in fields as diverse as Physics, Medicine, Psychology, Economics or Engineering. Among methods based on nonlinear time series analysis, recurrence plots and their quantification analysis are worthy of note.

The technique of recurrence plots was proposed by Eckmann et al. [1] as a graphical tool to study the time dependence behavior of a time series without any a priori assumption on the generator system. In order to quantify the most important features of a recurrence plot, Zbilut and Webber [2, 3], and Marwan [4] have, over the last ten years, developed the Recurrence Quantification Analysis (RQA).

The measures of RQA have been used by some researchers to detect bifurcation points in time series generated from systems whose behavior is well-known, like the logistic equation or the Lorenz system [5, 6]. The indicators introduced by Zbilut and Webber are able to find transitions between chaos and periodical states, and those developed by Marwan also identify chaos-chaos transitions [7]. We use recurrence quantification analysis to ascertain values of a control parameter that cause changes in the asymptotic behavior of a new dynamical system proposed recently by Liu et al. [8]. In their paper, the authors study the limit behavior of the system for different values of the control parameter in order to clarify the forming mechanism of the compound structure of the chaotic attractor existing when the control parameter is zero.

Here, we generate a time series for this model by increasing the control parameter in each iteration and we apply RQA to these data. This is a time series in which the transient behavior has not been removed, as is usual in this kind of applications [5, 6], in order to reduce the number of calculations. Other studies indicate that the performance of the RQA measures is quite similar to when a separate time series is generated for each value of the control parameter for which transients are removed [7].

In the following section, we review the recurrence measures which will be applied to a time series from Liu’s system in Section 4. In Section 3, we show the rich variety of complex dynamical behaviors of the system studied together with its Lyapunov exponents. Section 5 concludes with some comments and ideas for future research.

2. Recurrence Quantification Analysis

If we have a one-dimensional time series, \( \{x_i\}_{i=1}^n \), from an unknown dynamical system, we can reconstruct a phase state trajectory equivalent to the original phase space trajectory by embedding the one-dimensional time series in an m-dimensional space, using the delay coordinates method. The Takens and Stark theorems [9, 10] guarantee the preservation of the topological structures of the original trajectory when an embedding dimension \( m \) and a lag \( \tau \) are properly chosen. The embedded points are

\[ x_i = (x_i, x_{i+\tau}, ..., x_{i+(m-1)\tau}) \]

with \( i = 1, 2, ..., N \), where \( N = n - (m - 1)\tau \).

Recurrence plots are based upon the distances between the \( m \)-dimensional reconstructed points. Specifically, a recurrence plot is a symmetrical array \( N \times N \), in which a black point is placed at coordinates \( (i, j) \) whenever the vector \( x_i = (x_i, x_{i+\tau}, ..., x_{i+(m-1)\tau}) \) is close to the vector \( x_j = (x_j, x_{j+\tau}, ..., x_{j+(m-1)\tau}) \). The closeness has to be within a fixed threshold distance \( \epsilon > 0 \). It is also possible to define asymmetrical RPs or even to use different colors to represent different distances [6].

If a time series is deterministic, there will be diagonal and vertical lines in its recurrence plot, because recurrent behavior is a property of dynamical systems.
When the motion is periodic, the diagonal lines will be longer than those in a chaotic phenomenon. However, if the time series is independently and identically distributed, these lines will only exist by chance, so they will be shorter and the plot will present homogeneity in the distribution of recurrent points.

The principal advantage of this technique is that it does not impose any restriction on data distribution, nor on length nor stationarity. In terms of stationarity, a recurrence plot which is paling away from the principal diagonal would correspond to a nonstationary time series.

In the early 90’s, Zbilut and Webber [2, 3] developed some measures based on diagonal lines in recurrence plots. Ten years later, Marwan [4] carried out the quantitative analysis of vertical lines. Only five of these measures are used in this paper. The first three are related to diagonal lines and the last two to vertical lines:

Percent of recurrences: This represents the proportion of recurrent points, understood as two points which are separated by less than a fixed threshold distance $\varepsilon$. So,

$$\%REC = \frac{100}{N^2} \sum_{i,j=1}^{N} R_{i,j},$$

where $R_{i,j} = \Theta (\varepsilon - \|x_i - x_j\|)$, with $\Theta(\cdot)$ the Heaviside function and $\|\cdot\|$ a selected norm.

This is a measure of the density of recurrent points, which corresponds to the definition of correlation sum and reveals the probability of occurrence of similar states. For the dynamics to be properly described, it is important to choose an appropriate value for $\varepsilon$. Too large a value would lead to detecting recurrent behavior where there is no such behavior, while too small a value would result in a quantification of the noise only. The minimum percentage of recurrence required to obtain meaningful RQA is recommended as 1 [11].

Percent of determinism: This gives the ratio of recurrent points that form diagonals to all recurrent points,

$$\%DET = \frac{100}{\sum_{i,j=1}^{N} P(l)} \sum_{i,j=1}^{N} \sum_{l=l_{\min}}^{N} R_{i,j},$$

where $P(l)$ is the number of diagonal lines with length $l$. Note that a diagonal line of length $l$ starts in $(i,j)$ when $R_{i+k,j+k} = 1$ for $k = 0, 1, ..., l-1$. The minimal length considered for diagonal lines, $l_{\min}$, would have to be more than 2. Processes with stochastic behavior cause no, or very short, diagonals whereas deterministic processes are related to longer diagonal lines and fewer isolated recurrent points. This quantity is, therefore, a measure of predictability.

Maximal length of diagonal lines: This is related to the largest positive Lyapunov exponent. Periodic systems have longer diagonal lines than chaotic ones.

Laminarity: This is the percent of recurrent points forming vertical lines in an RP,

$$LAM = \frac{100}{\sum_{i,j=1}^{N} P(v)} \sum_{i,j=1}^{N} R_{i,j},$$

where $v_{\min}$ is the minimal length of vertical lines considered and $P(v)$ denotes the number of vertical lines with length $v$. A vertical line occurs when a state does not change or changes slowly and it has length $v$ starting at $x_i$ when $R_{i+j+k} = 1$ for $k = 0, 1, ..., l-1$. Usually $v_{\min} = 2$ is taken as the value to study transitions in the dynamics due to changes in the parameters of the system.

Trapping time: This refers to the average length of vertical lines,

$$TT = \frac{100}{\sum_{i,j=1}^{N} P(v)} \sum_{i,j=1}^{N} v P(v),$$

and it measures the mean time that the system is trapped in a specific state.

3. A new chaotic attractor: Liu’s attractor

Several three-dimensional quadratic autonomous systems, capable of generating different chaotic behavior, have been studied recently [12, 13]. One of them, proposed by Liu et al. [8], is described by the following nonlinear differential equations,

$$\begin{align*}
\dot{x} &= 10(y - x) \\
\dot{y} &= 40x - xz \\
\dot{z} &= -2.5z + 4x^2.
\end{align*}$$

(1)

For deeper knowledge of the forming mechanism of its chaotic attractor, these authors study the controlled system resulting from adding a parameter $e$ to the second equation. Using numerical simulation, Liu et al. [8] explore the behavior of the system for different values of $e$. When $e = 0$, the system has a two-scroll chaotic attractor with a compound structure, which they try to reconstruct by visualizing the shape of the attractor when $|e| = 14, 18$ and $40$. They obtained periodic structure when $|e| = 40$, a strange attractor with a single scroll when $|e| = 18$ and a strange attractor with two scrolls if $e = 0$. In all cases, the initial condition was $(2.2, 2.4, 38)$. 

372
In order to have a broader view of the behavior of the system, we calculate the largest Lyapunov exponent for $0 \leq e \leq 70$ using the same initial condition as [8]. Lyapunov exponents measure the divergence or convergence of nearby trajectories. A negative maximal Lyapunov exponent reflects that two trajectories approach a stable fixed point exponentially. If the motion settles down onto a limit cycle, the maximal Lyapunov exponent is zero. And if two trajectories diverge exponentially, the maximal Lyapunov exponent is positive, which indicates the existence of chaos. For the computation of the largest Lyapunov exponent, we have used the well-known Wolf algorithm [14]. The results are shown in Figure 1. There exists a large range of values for $e$ (from 0 to 33.5, approximately) corresponding to a chaotic motion, although with many short periodic windows. In another interval (from 33.5 to 229.13), there exists a limit cycle, and finally (from 229.13 to infinity) the attractor is a fixed point. Note that the system is dissipative for all values of $e$ (the divergence was always negative and equal to $-12.5$). Moreover, the graph is symmetric for negative values of $e$, so these have not been represented.

4. Application

In this section, RQA has been applied to find transitions in the asymptotic behavior of Liu’s system when the control parameter changes.

Firstly, as Trulla et al. or Iwanski and Bradley did with the logistic equation and the Lorenz system [5, 6], a time series was generated without removing transients. We numerically integrated Liu’s equations using a fourth-order Runge-Kutta method, increasing the parameter $e$ at each iteration step from 0 to 70, and we obtained a total of 150,000 three-dimensional points. The initial condition selected was $(2.2, 2.4, 38)$. In the subsequent analysis, the one-dimensional time series used was $\{x_i\}$, although similar results were obtained with components $y$ or $z$.

Before applying RQA, a state space trajectory equivalent to that of (1) was reconstructed from the one-dimensional series $\{x_i\}$. This requires choosing suitable dimension embedding, $m$, and delay, $\tau$. In this setting of nonstationarity, a sufficiently large embedding $m$ would have to be chosen such that the $m$-dimensional space would contain the relevant dynamics of the system from which the time series comes [11]. Since in our case the true system is known, we use $m = 7$ [9, 10]. Otherwise, standard methods should be employed, e.g. false nearest neighbors [15] or Cao’s method [16]. The minimum of the mutual information function was employed to calculate the delay $\tau$ [4]. Thus, with $m = 7$ and $\tau = 9$ a 7-dimensional time series of 149,964 data was obtained.

RQA was repeatedly performed on windows with 1,000 data from this 7-dimensional time series, verifying that two consecutive windows had 990 data in common. Thus, 14,895 values for each RQA measure were calculated. These measures have been computed using the RQA program developed by Zbilut and Webber.

In Figure 2(a), where determinism versus parameter $e$ is plotted shows three clear relative maxima for values of $e$ around 5.8, 11 and 20. This would indicate the existence of limit cycles near these parameter values. It is also observed that this indicator in general increases as of $e = 28.5$, with higher values than before. The high values near to 33.5 are also worthy of note. These results are confirmed by those obtained by the maximum length of diagonal lines (Figure 2(b)). In relation to the graph that shows the laminarity (Figure 2(c)), it is important to highlight that this measure decreases abruptly around $e = 16.3$. Moreover, the laminarity begins to decrease gently near $e = 33.5$, when periodic states start. Analogous behavior presents the trapping time statistics (Figure 2(d)), although this index also reveals a drop around $e = 28.5$, where another periodic window exists.
To summarize, it seems that the system presents the following different limit behaviors: for values of the parameter above 50, it has simple periodic behavior; when the parameter is between 33.5 and 50, approximately, the motion is cyclical but more complicated (period-doubling bifurcations); between 28.5 and 33.5 the behavior is chaotic, as for values smaller than 28.1. For \( e < 28.1 \) the attractor is chaotic but an increase in its complexity is detected for values below 16.3 (the chaotic attractor has two scrolls there, but only one for values greater than 16.3), although some cyclic windows with greater levels of determinism exist.

5. Conclusions

Measures based on the information contained in recurrence plots have been used in literature to detect points of bifurcation from time series generated from systems whose behavior is well-known, e.g. Lorenz system or logistic equation. This paper shows that some of these measures are effective in identifying values of the control parameter \( e \) around which the asymptotic behavior of Liu’s system changes. Specifically, the measures based on diagonal lines are useful to locate changes between periodic and chaotic behavior, and those constructed from vertical lines are better for distinguishing chaotic attractors of different shapes.

The analysis has been carried out using a time series with a specific length (150,000 data), but other simulations with smaller lengths have also been developed (75,000 and 50,000 points) recovering the same interval for the parameter \( (0 \leq e \leq 70) \), and we have obtained similar results, although these are not so clearly defined. A more detailed study would be necessary to identify exactly the information lost on account of the reduction in the length.

Further work needs to be done to theoretically understand the changes in the limit behavior of this system to different values of the control parameter \( e \). Furthermore, it would be convenient to study how important the choice of the step is in the integration method used to obtain the time series and to understand how the recurrence quantification indicators behave when the series is perturbed by noise.

References


Entropy Enhancement in a Chaos-Based Random Bit Generator

T. Addabbo†, M. Alioto†, A. Fort†, S. Rocchi† and V. Vignoli‡

† Dept. of Information Engineering, University of Siena
Via Roma, 56 - Siena 53100, Italy
Email: vignoli@dii.unisi.it

‡ Dept. of Electrical and Electronics Engineering, University of Bologna
Corso le Porte S. Giovanni, 1 – 40126 - Bologna - ITALY

Abstract—In this paper the theoretical bases to design a true random bit generator circuit with a predefined minimum entropy are discussed. The approach is tailored to generators based on the Sawtooth chaotic map, and it is based on a feedback control system designed for the optimization of the system parameters. The design approach was validated by an hardware prototype implemented on a Field Programmable Analog Array.

1. Introduction

An ideal True Random Bit Generator (TRBG) is a discrete time memoryless binary source with unbiased symbols that presents a Shannon Entropy of 1 bit/time-step. In practical TRBG circuit implementations the entropy, which is a quantity strictly dependent on the unpredictability and the randomness degree of the generated sequences, is lower than 1 bit/time-step. Since the entropy level can not be increased by means of deterministic post-processing algorithms without decimation or compression (i.e. without lowering the TRBG throughput), design approaches that achieve an adequate entropy at full throughput are of practical interest [1]. In most of the chaos-based TRBGs (see e.g. [1]-[3]), the theoretical knowledge about the chaotic system information generation mechanism allows for setting the nominal entropy arbitrarily close to the ideal maximum Shannon limit [4]. Nevertheless, the behavior of a chaotic circuit is in general highly sensitive to the tolerances of the implementation process, and therefore the actual entropy of the TRBG circuit often results lower than its designed nominal value [4]. In this paper, a feedback strategy to enhance the entropy of TRBG circuits based on the Sawtooth map that preserves the circuit full throughput is proposed [1], [4].

2. Chaotic TRBG circuit model

Let us consider the dynamical system (Sawtooth map)

\[ x_{n+1} = \begin{cases} 
    Bx_n + A, & \text{if } x_n < 0 \\
    Bx_n - A, & \text{if } x_n \geq 0,
\end{cases} \]  

where \( x \in \mathbb{R}, A, B \in \mathbb{R}^+ \). As discussed in [2], a possible implementation of (1) requires five different ideal analog blocks: one comparator which evaluates the sign of \( x_n \), one gain block \( B \), one adder, one \( \pm A \) constant generator (driven by the comparator digital output \( S_n \)), and one delay block. In [1], [4] it is shown that a random bit sequence \( b_n \) can be obtained by just picking out the binary sequence defined by the comparator output variable \( S_n \), that assumes the logic value ‘1’ if \( x_n > 0 \), and ‘0’ otherwise. Since parameter \( A \) just represents a scaling factor, the stochastic properties of the system generating \( b_n \), hereafter indicated as TRBG, are dependent only on the parameter \( B \), which must assume values greater than 1 for the system to be chaotic and not greater than 2 to avoid the state \( x \) being attracted to \( \pm \infty \). For \( B \in [\sqrt{2}, 2] \) the two unstable fixed points \( \pi_1 = A/(1 - B) \) and \( \pi_2 = A/(B - 1) \) define an interval \( I \subset \mathbb{R} \) such that any initial state \( x_0 \in I \) triggers a sequence \( x_n \) eventually attracted into the chaotic attractor \( A = [-A, A) \subseteq I \). Under these hypotheses, different binary sequences are related to different initial conditions, and their generation probabilities are related to the state invariant probability density function (pdf) of the chaotic dynamical system [2],[1].

When \( B = 2 \) the state system pdf in (1) approaches a stationary invariant uniform distribution over \( [-A, A] \). The amount of information supplied by a TRBG and its grade of redundancy are expressed by the Average Shannon Entropy (ASE) defined in what follows. Let be \( O_n = \{b_0, \ldots, b_n\} \) an n-bit length generated sequence, where \( b_n \in \{0, 1\} \). If \( P(O_n) \) is the generation probability of the sequence \( O_n \), the ASE in bit/time-steps is given by

\[ \text{ASE} = \lim_{n \to \infty} \left( -\frac{1}{n} \sum_{k=0}^{n} P(O_n) \log_2 P(O_n) \right) \]

where the summation extends over all the generable sequences and the limit is approached from above [2]. Kocarev and Stojanovski showed in [4] that the ASE of the TRBG is an increasing function of the parameter \( B \), and that it reaches the upper ASE limit of 1 bit/time-step for \( B = 2 \). This result focuses the trade-off imposed to the TRBG circuit designer: to avoid the \( B \) value being greater than 2 (also accounting for the actual circuit tolerances) while trying to achieve a \( B \) value as close as possible to this limit in order to maximize the ASE.

2.1. Modeling the non-idealities

In Fig. 1 a modified block diagram that accounts for the most important non-idealities of circuits implementing (1) is shown, and in (3) the corresponding modified map is...
Equation (3) can be rearranged as

\[ x_{n+1} = \begin{cases} 
Bx_n + W + Z_B, & \text{if } x_n < P \\
Bx_n - W + Z_B, & \text{if } x_n \geq P,
\end{cases} \quad (4) \]

where \( W = A', Z_B = B \cdot OS_B + OS_A \), and \( P = OS_C \). The parameters \( (W, Z_B, P) \) determine the map scale factor and the map position in the plane \((x_n, x_{n+1})\), while the parameter \( B \) sets the slope of the map (Fig.2).

### 2.2. Equivalent dynamical systems

In this paper two different dynamical systems (4) are said equivalent if they define two TRBGs characterized by the same stochastic properties (i.e. for each binary sequence \( O_n \) the generation probability \( P(O_n) \) is equal for both of the TRBGs). From this point of view, it has been proved by the authors in [2] that two systems (4) with a same \( B \) value are equivalent (and therefore the statistical properties of the output bit sequences \( [b_n] \) are equal) if they have the same central distance \( D \), where \( D \) is defined as

\[ D = \frac{P(B - 1) + Z_B}{W \sqrt{2}}. \quad (5) \]

Referring to Fig.2, the absolute value of \( D \) is equal to the Euclidean distance of the point \( C = (P, BP + Z_B) \), in the plane \((x_n, x_{n+1})\), from the bisector of the first and third quadrants, normalized to \( W \). Accordingly, for any given \( B \) value, systems described by (4) are equivalent to the following equivalence class representative dynamical system:

\[ x_{n+1} = \begin{cases} 
Bx_n + 1 + D \sqrt{2}, & \text{if } x_n < 0 \\
Bx_n - 1 + D \sqrt{2}, & \text{if } x_n \geq 0.
\end{cases} \quad (6) \]

Moreover, given \( B \), systems (1) and (4) are equivalent (and therefore the corresponding TRBGs have the same ASE) if the central distance \( D \) of (4) is zero, that is if

\[ P(B - 1) = -Z_B. \quad (7) \]

### 3. Enhancing the ASE: map correction

When the central distance \( D \) is not zero, to avoid the state \( x \) of (4) to jump beyond the fixed points \( \pi'_1 \) and \( \pi'_2 \) (Fig.2) and to be attracted toward infinity (with a consequent electronic saturation in practical implementations), \( B \) must satisfy the relationship

\[ 1 < B \leq \frac{2}{1 + |D| \sqrt{2}} \leq 2. \quad (8) \]

According to (8), the quantity \( |D| \) determines an upper bound for the parameter \( B \) which fixes the rate of mixing of the chaotic map [1], and if \( D \) tends to zero, the maximum allowed \( B \) value tends to 2. Therefore, as discussed in [2], a non-zero central distance \( D \) determines also an upper bound for the ASE achievable by the TRBG. This point has
been investigated by the authors on the basis of the computation of the asymptotic invariant pdf of the dynamical system (6) (Fig. 3).

The ASE maximization procedure proposed in this paper is based on the fact that the two parameters $D$ and $B$ are sufficient for determining the stochastic properties of the TRBG. First of all, from (7), if $B$ is kept constant, $D$ can be zeroed by changing either the parameter $P$ or the parameter $Z_B$, that is, referring to Fig.1, by injecting just one correction offset either in the adder $\Sigma_1$ or in the adder $\Sigma_2$. Once minimized $D$ with a further correction $B$ can be increased according to (8).

In practical systems the required exact values of the correction signal for $D$ and $B$ are not available, since the quantities $OS_A$, $OS_B$, and $OS_C$ are not known. Nevertheless, the values of the correction signals for $D$ and $B$ can be obtained from the statistical analysis of the generated binary sequences, as it was discussed by the authors in [2]. In particular, if $P(b_n = 1)$ is the probability for the $n$-th binary output to take the ‘1’ logic value, and if $\mu = (P(b_n = 1) - 0.5)$ for $n \to +\infty$, the authors showed that for each $B$ value the parameter $\mu$ is a monotonic function of $D$, and that $\mu = 0$ when $D = 0$. As a consequence, the sign estimation of $\mu$ can be used to minimize $|D|$ [2]. Anyway, since the upper ASE limit of 1 bit/time-step is reached when $B = 2$ (Fig. 3), and due to the upper limit for $B$ given by (8), the ASE is maximized if both $D \to 0$ and $B \to 2$. Summarizing, the simplest way to drive system (4) to a desired operating point $(B_{\text{max}},|D|_{\text{lim}})$, where $B_{\text{max}}$ is the maximum value of $B$ according to (8) given $|D|_{\text{lim}}$ is to iteratively minimize the central distance magnitude $|D|$, exploiting the estimation of the $\mu$ sign, and then increase $B$ until the divergence of the state orbit is detected [2].

4. Correction signals resolution and worst case ASE

As discussed in the previous section, two correction signals $S_B$ and $S_D$ are sufficient for correcting the map: the former signal is used for changing the amplification $B$ of the gain block in Fig.1, whereas the signal $S_D$ is used to minimize $D$. Even if the correction technique described above may be used to drive system (4) to any desired operating point $(B_{\text{max}},|D|_{\text{lim}})$ arbitrarily close to the optimum point $(2, 0)$, in practical realization the precision of the correction system is limited by the physical resolution of signals $S_D$ and $S_B$. Accordingly, by denoting with $\delta_B$ and $\delta_D$ the minimum variations effectively achievable on gain $B$ and on signal $S_D$ respectively, a worst case operating point $(B_{\text{max}},|D|_{\text{lim}, wc})$ can be estimated as a function of $\delta_B$ and $\delta_D$. In particular, when injecting the correction signal $S_D$ in node $\Sigma_1$ (a similar analysis can be carried out considering the injection of $S_D$ in node $\Sigma_2$), the map (3) must be modified as

$$f'(x_n) = \begin{cases} B (x_n + OS_B) + A' + OS_A, & \text{if } x_n < OS_C + S_D \\ B (x_n + OS_B) - A' + OS_A, & \text{if } x_n \geq OS_C + S_D. \end{cases}$$

(9)

According to (5), since $D$ for the system (9) is equal to

$$D = \frac{OS_C(B-1) + OS_A + B \cdot OS_B + S_D(B-1)}{A \sqrt{2}},$$

(10)

$D$ is zeroed if

$$S_D = \frac{OS_C(B-1) + OS_A + B \cdot OS_B}{1 - B},$$

(11)

In the worst case it can be assumed that, after the $|D|$ minimization, the $S_D$ actual value differs from (11) by the quantity $\delta_D/2$, and we can write

$$|D|_{\text{lim}} \leq \frac{(B-1)\delta_D}{2A'\sqrt{2}} = D_0.$$  

(12)

If, once minimized $|D|$, also the $B$ value is changed by $S_B$, the value of $D$ changes due to (5). In detail, if $B$ is increased by $\delta_B$, (12) is modified according to

$$|D|_{\text{lim}} \leq D_0 + \frac{\delta_B}{A'\sqrt{2}} \left( \frac{|OS_B|_{\text{max}} + |OS_A|_{\text{max}}}{B-1} + \frac{\delta_D}{2} \right).$$

(13)

The upperbounds expressed in (8) and (13) represent the worst case limit for the control system, and in practical realizations, even if performing the correction, a worst case operating point $(B_{\text{max}} < 2, |D|_{\text{lim}, wc})$ must be taken into account. This limit, that depends on the resolution of the correction signals, identifies the minimal guaranteed ASE for the corrected TRBG.

5. Experimental results and conclusions

The proposed correction procedure, described in detail in [2], was tested on a prototype TRBG circuit implemented by a Field Programmable Analog Array (FPAA Anadigm AN212E04). This SRAM based device can be dynamically reconfigured by an host processor while the old configuration is still active and running. The activation
of the new configuration happens in real time, allowing for an easy implementation of the proposed ASE maximization strategy. The gain block $B$ in Fig. 1 was implemented using a FPAA library adjustable gain block, whose profile of gain versus control voltage was specified to linearly vary within the nominal range $1.4 \div 2.08$ with an experimentally verified resolution $\delta_B \approx 2 \cdot 10^{-2}$ for the working frequency of 250 kHz. The control and correction procedure for the ASE maximization was software implemented on a PC equipped with an acquisition board (National PCI MIO64E) for acquiring the TRBG output, and whose 12 bit D/A converters, with a proper scaling, provided the control signals $S_B$ and $S_D$. The correction system was tested using an experimentally verified resolution $\delta_c \approx 5 \cdot 10^{-3}$, while setting for the map parameter $A$ in (1) an equivalent voltage nominal value of 1.8V. With the aim of emulating different combined effect of $OS_A$, $OS_B$ and $OS_C$ (that depend on the specific circuit implementing the map), exploiting the finest resolution of the PCI MIO64E device at disposal, several initial operative conditions were forced to the prototype by adding random offsets to the correction signals [2]. Partial ASE values as high as 0.99 bit/time-step (for word lengths up to 16 bit) were experimentally achieved at the maximum allowed speed of 250 kHz, without performing any kind of digital post-processing on the output bit sequences. This result agrees with the theoretical worst case ASE obtainable considering Fig. 3 and the upper bounds (8) and (13), evaluated referring to the typical value of $OS_A$, $OS_B$ and $OS_C$ obtainable from the FPAA data sheet [2]. In Fig. 4 and 5 the effect of the correction procedure on the actually implemented map is shown. In each tested case, it has been estimated that after the correction the prototype was working with $B$ values not lower than 1.99, obtaining an invariant distribution of the state over the chaotic attractor very close to the ideal uniform distribution theoretically achievable for $B = 2$. These results were obtained with resolutions of the correction signals $S_B$ and $S_D$ lower than the actual resolution normally achievable in accurate-designed integrated analog circuits.

Without post-processing, the proposed generator was able to issue sequences passing the FIPS 140.2 tests with a success rate higher than 98%, while typically failing only 4 of 18 Diehard tests [2]. On the other hand, with the direct bit-by-bit XOR with the output of a 8 bit Linear Feedback Shift Register, the sequences of the proposed TRBG passes all of the considered tests at full throughput.

The effectiveness of the proposed technique, proved from a theoretical point of view, is not tied to the specific implementation presented in this paper, and can be successfully applied also to integrated designs with an increase in the output bit rate of up to two order of magnitude, with state of the art CMOS technologies, with respect to the FPAA implementation.

References


A Study on Generation of Aperiodic Random Numbers
Using Tent-Map-Type Feedback Shift Registers

Akio Tsuneda, Takashi Tanoue, Sho Mitsuishi, Daisaburo Yoshioka, and Takahiro Inoue

Department of Computer Science and Electrical Engineering, Kumamoto University
2–39–1 Kurokami, Kumamoto, 860–8555 JAPAN
E-mail: tsuneda@cs.kumamoto-u.ac.jp

Abstract—There are several attempts to use chaotic sequences generated by nonlinear maps. Some chaotic maps can produce balanced and i.i.d. (independent and identically distributed) binary sequences. However, it is difficult to generate aperiodic sequences with good statistical properties by analog circuits or digital circuits only. In this paper, we propose a new sequence generator consisting of aperiodic sequence generator and digital feedback shift registers. Some numerical results are shown.

1. Introduction

Random number generation is one of important technologies for several engineering applications such as Monte Carlo simulations, cryptosystems, spread spectrum CDMA communications. Especially, Monte Carlo simulations need random numbers with very long period [1], preferably with infinite period such as physical random numbers.

One of methods to obtain aperiodic sequences is to use chaos which is defined as random phenomena generated by simple deterministic systems and some of them can be theoretically analyzed [2],[3]. For example, by using one-dimensional (1-D) discrete-time chaotic systems, we can theoretically design truly random sequences [4],[5]. If we realize such chaotic systems by analog circuits, aperiodic random sequences can be generated. Note that digital circuits generates only (eventually) periodic sequences. However, it is very difficult to generate random sequences with good statistical properties due to nonidealities of analog circuit elements and inevitable noise [3],[5]. We have been trying to improve the statistical properties of aperiodic binary sequences generated by analog chaos circuits [6],[7].

In this paper, we propose a simple method to generate aperiodic random numbers with good statistical properties based on a kind of feedback shift register (FSR) which realizes the tent map with finite bits. Such discretized tent maps have also been considered in [8]–[10] for encryption, spread-spectrum communications and coded modulations. In addition to FSRs realizing the discretized tent map, we assume that we have a generator of aperiodic random numbers (e.g., an analog chaos circuit) which is just used to guarantee the aperiodicity.

Furthermore, we improve the proposed generator by cascading two proposed ones. By numerical experiments, we investigate some statistical properties of random sequences generated by the proposed generator.

2. One-Dimensional Map and Chaos

One-dimensional (1-D) nonlinear difference equation

\[ x_{n+1} = \tau(x_n), \quad x_n \in I = [d, e], \quad n = 0, 1, 2, \ldots \]  

(1)
can produce a chaotic real-valued sequence \( \{\tau^n(x)\}_{n=0}^{\infty} \) starting from an initial value \( x_0 = x \), where \( \tau^n \) denotes the \( n \)-th iteration of the map \( \tau \). For a chaotic sequence \( \{G(\tau^n(x))\}_{n=0}^{\infty} \), its average is given by

\[ E[G] = \int G(x)f^*(x)dx, \]  

(2)

where \( f^*(x) \) is the invariant density for the map and important for evaluation of statistics of chaos [2],[4].

Define a threshold function by

\[ \Theta_t(x) = \begin{cases} 0 & (x < t) \\ 1 & (x \geq t) \end{cases} \]  

(3)

Then we can obtain a chaotic binary sequence \( \{\Theta_t(\tau^n(x))\}_{n=0}^{\infty} \). We can also use a binary function given by

\[ B(x) = \sum_j (-1)^j \Theta_t(x), \]  

(4)

which is a combination of threshold functions. If we use a class of chaotic maps and appropriate binary functions \( B(x) \), we can theoretically design balanced and i.i.d. (independent and identically distributed) binary sequences which are ideal random numbers [4],[5].

3. Random Number Generation Based on Tent Map

3.1. Proposed Generator

Figure 1 shows the proposed random number generator which consists of a shift register with \( k \) memory cells and an aperiodic random bit generator, where \( \oplus \) denotes modulo-2 addition. At each clock, the value (0 or 1) of each memory cell is shifted to the right after modulo-2 added with \( a_0(n) \), and \( a_{k-1}(n) \) is updated by \( a_{k-1}(n + 1) = \).

379
aperiodic random bit generator
(e.g., analog chaos circuit)
: modulo-2 addition
(XOR)

\[ a_{k-1}(n) \oplus b_k, \] where \( b_k \) the output of the aperiodic random bit generator. The state of the register at time \( n \), denoted by \( \{a_{k-1}(n), a_{k-2}(n), \ldots, a_1(n), a_0(n)\} \), is converted into a real number \( x_n \in [0, 1] \) by

\[ x_n = a_0(n) \cdot 2^{-1} + a_1(n) \cdot 2^{-2} + \cdots + a_{k-1}(n) \cdot 2^{-k}. \quad (5) \]

Plotting \( (x_n, x_{n+1}) \), we can get 1-D maps (so-called return maps) as shown in Fig.2, where (a) \( k = 5 \) and (b) \( k = 10 \). This result shows that the shapes of such return maps are similar to the tent map defined by

\[ \tau_T(x) = \begin{cases} 2x & (0 \leq x < \frac{1}{2}) \\ 2(1 - x) & (\frac{1}{2} \leq x \leq 1) \end{cases} \quad (6) \]

which is one of well-known chaotic maps. Thus, this generator is considered to approximate the tent map with finite bits.

Note that a real-valued sequence generated by the original tent map is theoretically shown to have a uniform probability density function and white spectrum. It should also be noted that a binary sequence \( \{\Theta_{\frac{1}{2}}(\hat{x}_n)\}_{n=0}^{\infty} \) obtained from a real-valued sequence \( \{\hat{x}_n\}_{n=0}^{\infty} \) generated by the original tent map is theoretically shown to be balanced (to have equiprobability of 0 and 1) and i.i.d. [4]. Since \( \Theta_{\frac{1}{2}}(\hat{x}_n) \) corresponds to \( a_0(n) \) in eq.(5), we can expect the output bit sequence of the proposed generator tends to be balanced and i.i.d. when \( k \) becomes large.

### 3.2. Statistical Properties

We performed some numerical experiments concerning statistical properties of sequences generated by the proposed generator. We already confirmed that the probability densities of real-valued sequences \( \{x_n\}_{n=0}^{\infty} \) obtained by eq.(5) are almost uniform, which is independent of the outputs of the aperiodic random bit generator [7].
Next, we investigated \( m \)-distributivity of binary sequences generated by the proposed generator. Here, a binary sequence \( \{ a_n \}_{n=0}^{\infty} \) is said to be \( m \)-distributed if

\[
\text{Prob}(a_n a_{n+1} \cdots a_{n+k-1} = B_1 B_2 \cdots B_m) = \frac{1}{2^m} \quad (7)
\]

for all binary numbers \( B_1 B_2 \cdots B_m \) [11]. Namely, we observed successive \( m \)-bit patterns in the sequences and enumerated the number of each pattern in \( 2^m \) patterns. Here we used imbalanced and correlated binary sequences as the inputs to the aperiodic random bit generator, where \( p = E[b_n] = 0.7 \) (probability of 1) and the correlation parameter \( a = 3 \) (auto-correlation function is given by \( a^{-l} \)).

Figures 3 (a) and (b) show 7-distribution and 8-distribution of binary sequences generated by the proposed generator with \( k = 7 \), where the horizontal axis indicates each \( m \)-bit pattern by decimal integers and the sequence length is \( 8000 \times 2^m \). We can find that the sequence is 7-distributed but not 8-distributed. We confirmed that the binary sequences are \( m \)-distributed if \( m \leq k \).

Furthermore, Fig.4 shows the autocorrelation function \( E[(a_0(n) - 0.5)(a_0(n+l) - 0.5)] \) of binary sequences generated by the proposed generator with \( k = 7 \), where \( p \) is the probability of 1 in the binary sequences and \( a \) is a correlation parameter. The strength of the correlation increases as \( a \to 1 \). From this figure, we find that the peak values of the autocorrelation appear when \( l = r(k+1) \) (\( r = 1, 2, \cdots \)). The peak values increase as \( a \to 1 \) or \( p \to 1 \). Thus we need to suppress such peaks [7].

4. Improvement of Proposed Generator

We cascade two proposed tent-map-type FSRs (feedback shift registers) as shown in Fig.5. The output binary sequences from the 1st FSR have statistical properties shown in the previous section, that is, they have some peaks of auto-correlation values but they are balanced and \( m \)-distributed for \( m \leq k \). Thus the 2nd FSR uses the 1st one as an aperiodic binary bit generator, where the number of stages of the 1st FSR is denoted by \( k \) and the 2nd one is denoted by \( r \). Figure 6 shows \( m \)-distributions of binary sequences generated by the improved generator. We find that they \( m \)-distributed if \( m \leq k + r \).

Next, Figure 7 shows auto-correlation properties of binary sequences generated by the improved generator. In the case of (a) \( k = 3 \) and \( r = 4 \), small peaks appear when \( l \) is a multiple of 20. We conjecture that peaks appear when \( l \) is the least common multiple of \( k + 1 \) and \( r + 1 \). However, it is shown from the case of (b) \( k = 5 \) and \( r = 6 \) that the first (maximum) peak can be reduced by using large \( k \) and \( r \). This is independent of statistical properties of aperiodic binary sequences inputted to the 1st FSR. Hence, the improved generator is useful for random number generation.

5. Conclusion

A random number generator using aperiodic random bit generator and feedback shift registers has been proposed. Binary sequences generated by a single proposed generator have some peaks of their auto-correlations. However, the statistical properties can be much improved by cascading two proposed generators if the length of the registers is large, which is independent of statistics of the original aperiodic binary sequences. Hence, we can obtain aperiodic binary sequences with good statistical properties by the proposed generator.

References

Figure 6: $m$-distribution ($k = 3, r = 4, p = 0.8, a = 2$)


Figure 7: Auto-correlation properties generated by the improved generator


Simple and Effective Post-Processing Stage for Random Stream Generated by a Chaos-Based RNG

Fabio Pareschi†, Riccardo Rovatti‡, and Gianluca Setti†

† ENDIF, University of Ferrara
Via Saragat, 1 - 44100 Ferrara, ITALY
‡ DEIS, University of Bologna
Viale risorgimento, 2 - 40133 Bologna, ITALY
All authors are also with ARCES - University of Bologna
Via Toffano, 2 - 40125 Bologna, ITALY
Email: fparesci@ing.unife.it, rrovatti@deis.unibo.it, gsetti@ing.unife.it

Abstract—We present design and validation of a prototype of a true random number generator which internally exploits a pipeline analog-to-digital converter modified to operate as a set of chaotic maps. The prototype has been implemented in AMS 0.35 µm 2P3M technology and has a nominal throughput of 40 Mbits/sec. The circuit can be formally proved to deliver perfectly uncorrelated output sequences. Here we take into account the effects of implementation errors by analyzing sequences generated by the prototype, and we introduce a data post-processing methodology suitable for their compensation and for assuring robust behavior, looking for the simplest architecture that ensures a certain quality of the generated sequences.

1. Introduction

Random number generators (RNGs) are gaining more and more interest, due to the increasing usage of cryptography, where RNGs represent a critical point [1]. Randomness is widely used in many cryptographic primitives; for instance it is fundamental in the synthesis of confidential keys for symmetric and public-key cryptosystems.

Traditionally the most used sources of randomness rely on pseudo-random number generators (PRNGs). These are algorithms capable of “expanding” short seeds into long, irregular bit sequences. The procedure is in fact non random, and the only source of entropy is actually represented by the seed.

The recent, huge growth of data-security applications is questioning the general application of PRNGs, since no cryptographic algorithm can be stronger than its underlying RNG [2]. All major IT players [3, 4] are now developing secure platforms based on true random number generators (TRNGs) as opposed to PRNGs. TRNGs are conventionally based on the explicit observation of analog quantities from complex physical processes such as Johnson thermal noise [3], oscillator jitter noise [4], turbulence or even quantum effects like the radioactive decay. Especially RNG based on quantum effects can achieve very good results in terms both of quality and speed (which are the two most important figures of merit in a RNG); recently, some RNGs based on photon reflection [5] has been proposed. However they do not represent a solution currently embeddable in silicon integrated technology. So if we look for a system-on-a-chip solution, or just for a high quality low-cost solution, we must rely on processes which could be easily reproducible in a standard silicon-based IC technology.

Here we refer to a TRNG based on a set of simple one-dimensional chaotic maps [6], that follows the approach recently proposed in [7], where a pipeline analog-to-digital converter (ADC) based on 1.5 bit/stage cells [8] has been reconfigured to implement a chaotic circuit that has been theoretically proved to generate independent and identically distributed - that is random – symbols. A prototype of the proposed TRNG has been designed and fabricated in AMS 0.35 µm 3.3 V CMOS technology.

Here we complete the circuit with a post-processing stage. This typically consists of a very simple algorithm which elaborates the output bit-stream in order to reduce the residual correlation, if still present. It is always used in TRNGs to improve the quality of the output stream [3, 4].

The paper is organized as follows. In section 2 we present the design of the circuit, while in section 3 we present two different post-processing techniques that we apply to the streamout of the prototype. Results for statistical tests on sequences generated by the prototype and post-processed by the proposed architecture are reported. For the latter we use the test suite provided by the US National Institute for Standard and Technology (NIST) [9].

2. Circuit description

The design of the proposed circuit comes directly from pipeline ADCs technology. The architecture of a standard pipeline ADC, including the proposed modifications, is shown in Fig. 1.

Practically, in [7] it is observed that a pipeline A/D converter closed into a loop after k stages, has the same behavior as k independent chaotic maps.

Generally speaking, chaotic maps are 1-D discrete-time autonomous systems, whose evolution is described as $x_{n+1} = M(x_n)$ where $M$ is a proper function mapping an interval $I$ into itself. A deep and exhaustive analysis on chaotic maps, as well as a description of the tools neces-
DAC—ADC pipeline composed by eight stages. The circuit is based on an analog state of the map \( x_n \) is reflected into a loop and re-elaborate the digital output of all stages, i.e. to substitute the digital correction logic, as shown in Fig. 1.

The microphotography of the designed circuit is shown in Fig. 2. Two pipelines have been implemented in this circuit; a two-stages pipeline including analog buffers whose purpose is to test the analog core, and an eight-stages to test the behavior of the circuit as a RNG. Here we refer to the pipeline composed by eight stages. The circuit is based on 1.5 bit A/D cell [8]; with this architecture it is possible to generate 1 bit/stage every time step. The nominal working frequency of the circuit is 5 MHz, so the throughput of the circuit is 40 Mbit/s.

Since the prototype is designed only for testing purpose, no post-processing stages have been integrated.

3. Post-processing stages and results

In this paper we consider sequences generated by the chaos-based RNG and test them for randomness after processing them. We consider two different kinds of post-processing.

The first one is a simple parity based post-processing (also know as xor post-processing). In this case the output of the system is computed as the parity of non-overlapping strings of \( n \) bit, resulting in a decimation of \( 1:n \).

Note that even if introducing decimation is an effective and easy way to improve the quality of a random stream, it could be not sufficient. In fact, if a small correlation is present between symbols more distant than \( n \), it will be still present after applying this post-processing. This means that decimate a random sequence is not always the better way to eliminate the residual correlation, while at the same time it is extremely expensive in terms of data rate reduction.

On the contrary, the second kind of post-processing that we propose in this paper is a constant bit-rate processing. Its basic structure is reported in Fig. 3. Strictly speaking, the working principle of the proposed stage is to elaborate bits coming from the RNG and memorize them into a shift-register and to reuse these bits to xor the same bit-stream coming from the RNG after few time steps. This basic operation is repeated few times to increase the efficiency.

This post-processing is composed by four shift register A, B, C, D, of lengths \( a, b, c \) and \( d \) respectively, which can be chosen arbitrarily; Note that, for their particular structure, the shift-register B and D has to be composed at least by two stages, while A and C can be constituted by a single flip-flop.

For testing these post-processing stages, we used the NIST SP 800-22 test suite [9] which is, to the best of the authors’ knowledge, the most reliable test suite currently available. The code we used for the testing comes directly from NIST website [10] and is the latest version available the moment (version 1.8). The NIST suite is composed of 15 tests. However, due to some known error on the reference distribution on the Spectral (DFT) test [11], only the
Each test has to be interpreted in a statistical way [9]; this means that after analyzing a sequence, we get only a real number in [0, 1], called P-value, that can be interpreted as the probability that the sequence has been generated by an ideal RNG. A P-value equal to 1 means that the sequence has the maximum probability to have been generated by an ideal generator, while a value near to 0 means that it is very unlikely, though still possible, that the sequence is effectively random.

Due to this reason, we have to be very careful when looking at randomness tests results. According to NIST, we can say that a test is passed if we get a P-value greater than a chosen significance level α, which NIST suggests to take equal to α = 0.01. However, if getting a P-value lower than α means that there is a very low probability (i.e. lower then α) that the sequence is effectively generated by an ideal RNG, getting a P-value greater than α does not imply that the sequence has been randomly generated.

In order to overcome the impasse, we consider a different approach, i.e. we perform a second level test. Instead of analyzing a single sequence, we acquire a large number of sequences, post-process and analyze them, and collect all the P-values. Then, since for a true-random sequence a P-value is a random variable uniformly distributed in [0, 1], we check the uniformity of P-values with a χ² test. The result of this test is again a real number in [0, 1] that can be considered as the probability that the samples are coming from a uniformly distributed process, i.e it can be considered as a second level P-value. As in the standard case, we can choose another significance level α′, and consider the test passed if this second level P-value is greater than α′. We can consider, as in the previous case, α′ = 0.01. Though this test still has to be interpreted in a probabilistic way, and still a truly RNG has a probability equal to α′ to not pass this test (type I error, see [9]), the probability that a non-random generator has to pass this test (type II error), which is a quantity completely unrelated to the previous one, is expected to be much lower than in the previous case. In other words, this test is much more reliable for detecting non-randomness than first level test.

Results are shown in Tables 1 and 2. In Table 1 we report the results for processing data coming from the prototype of the RNG running at the nominal speed of 5 MHz, i.e. with a throughput of 40 Mbit/s, while in Table 1 we report the results from the prototype overclocked at 6 MHz, i.e. with a throughput of 48 Mbit/s. We analyzed 10,000 sequences of length (after the decimation introduced by the post-processing) equal to 1 Mbit and computed the χ² value over 16 bins.

In term of post-processing, we use the following parameters:

- **Xor based post-processing**: this post-processing has been considered with depth equal to 4, 8 and 16 bit, i.e. with a decimation rate equal to 1:4, 1:8 and 1:16.

- **Shift-register based post-processing**: we have considered 5 possible architectures with different values of the four parameters a, b, c and d, corresponding to a complexity of the stage from 6 to 14 flip-flops.

In both case we can notice that increasing the complexity of the system is reflected in a better yield of the tests (i.e. in a better quality of the random stream). In the first case increasing the complexity means increasing the length of the string of which we compute the parity bit; this implies an increment of the decimation rate, i.e. a reduction of the speed of the circuit. The effective speed of the bitstream after the post-processing ranges from 10 to 2.5 Mbit/s at the nominal speed and from 12 to 3 Mbit/s for the overclocked circuit. We can notice that while a depth of 4 is enough for passing all tests when the circuit is working at the nominal speed, we have to increase the depth up to 16 for passing all tests at the overclocked speed of 6 MHz.

In the second case, increasing the complexity means increasing the number of flip-flops, i.e. increasing the hardware cost (both area and power consumption). However, due to the particular structure, no decimation is introduced, so we are effectively analyzing a bit stream generated at the speed of 40 and of 48 Mbit/s. For passing all tests at the nominal speed, we need a minimum complexity of 10 FFs. For the circuit overclocked at 6 MHz, the minimum required complexity is 12 FFs.

4. Conclusions

We have tested for randomness a prototype of a chaos-based RNG designed in 0.35 µm CMOS technology, able to generate up to 40 Mbit/sec, and overloaded to reach a throughput of 48 Mbit/sec, applying two different very simple post-processing stages.

Furthermore, the testing phase has not involved a single sequence but a large number of them, providing in this way results much more reliable. For both post-processing we have found a minimum complexity of the system that allow us to consider, both for the system working at the nominal speed and for the system overclocked, the chaos-based RNG as a high quality TRNG, suitable for the most advanced security-related applications.

The other figure of merit in a RNG is the throughput. The shift-register based post-processing introduces no decimation, so the throughput of the system including the post-processing is still 40 Mbit/sec (or 48 Mbit/sec for the overclocked circuit). This is a considerably high speed, especially when compared the the other generators available. A
high-end quantic RNG like the one described in [5], which is one of the fastest available, achieves a throughput of 4 Mbit/s with a much higher power consumption. If we instead compare our performance with other silicon-based architecture, the gap is very large. For example the Intel TRNG [3], which is based on Johnson thermal noise and on Von Neumann post processing, achieves only an average speed of 75 Kbit/sec. VIA declares [4] that its generator can achieve a data-rate from 4 to 9 Mbits/sec, but with a much smaller entropy rate. However, no results from testing is presented in both cases.

References


On the Robustness to Noise and Interference of ADC-Derived, Chaos-Based True Random Number Generators

Sergio Callegari

ARCES and DEIS, University of Bologna
Via Toffano 2/2, Bologna, Italy
Email: scallegari@arces.unibo.it

Abstract — Recent results in the design of true-Random Number Generators (RNGs) include a technique for building true-RNGs out of Analog to Digital Converter (ADC) parts, where chaotic dynamics is exploited to guarantee appropriate statistical features and unpredictability. A major advantage of this approach consists in the possibility to formally prove the ability of the system to generate true-random bits in ideal conditions (building blocks without deviations from nominal behavior and noise/interference-free environment). Here, the original theoretical framework is extended to take into account noise and interference, proving that a system made of ideal building blocks is insensitive to them.

1. Introduction

Random Number Generators are required in a large number of engineering applications and true-RNGs are specifically needed in crypto-systems [1], where random numbers are inherent in the computation of algorithms such as the DSA, in the synthesis of confidential keys for symmetric- and public-key crypto-systems as RSA moduli, and in many communication protocols [2].

Unfortunately, the deployment of true-RNGs has traditionally been hindered by high design costs. As a matter of fact, true-RNGs are conventionally designed as physical-RNGs, i.e. as systems where the macroscopic effect of some microcosmic phenomenon is observed: it is the complexity of the latter to make the macroscopic observable random by all practical means. Examples are sources based on thermal noise, quantum noise, emission timing in radioactive decay, and so on. The economic and practical issues are self evident: not only physical-RNGs require analog design, but also the most unsystematic kind of it. Indeed, most of the abstractions that usually insulate the designer from the complexities of hardware must be abandoned to delve into the details of the underlying technology. Furthermore the final resilience to external interference remains hard to determine and questionable.

Interestingly, in spite of the above problems the importance of true-RNGs is such that many large manufacturers have brought physical-RNGs to the market [3, 4]. With this, the recent revelation that simple chaotic circuits can be used to build true-RNGs without the need to directly observe physical quantities [5, 6] has clearly been most welcome. The major advantage of the chaos-based approach is to make designs more structured thanks to the possibility of relying on deterministic models based on signal operations that easily map into standard circuit primitives (adders, differentiators, gain blocks, etc.). Consequently, even if the chaotic approach cannot free the designer from analog circuits, at least it can largely reduce their cost. Notably, one of the chaotic-RNGs proposed so far comes to the point of completely re-using the analog signal processing chain of pipeline ADCs [7], also opening the way to reconfigurable ADC/RNG architectures [8].

As a second advantage, some chaotic circuits (and notably for those based on Piece Wise Aline Markov (PWAM) maps [9]) allow mathematical toolboxes to be used to formally prove the ability to operate as true-RNGs. Clearly, this is a fundamental asset since no empirical test can be as general and convincing as a formal proof. Unfortunately, there is currently a limitation in that mathematical proofs can only be derived with regard to ideal conditions, namely for ideal circuits placed in noise/interference-free environments [7]. This is due to the inherent difficulties in the analysis of perturbed chaotic systems.

In this paper, the formal analysis of the ADC-derived chaotic-RNG presented in [7] is extended to consider the effects of noise and interference. So far only intuitive arguments were used to suggest that limited amounts of noise could even be beneficial [7, Section 4] and nothing at all was said about interference. Here, formal arguments are employed for the first time to prove that chaos based RNGs can be completely insensitive to noise and interference when their building blocks conform to their nominal behavior.

2. Background

Figure 1 shows the general architecture of an ADC based true-RNG, composed by an arbitrary, even number of stages, each of which implements $(1 + \frac{1}{2})$-bit A/D conversion and vice versa. Note that the requirement for an even number of stages uniquely comes from the Sample and Hold (S/H) elements. A single stage system can be devised by introducing an additional S/H as shown in figure 2. For the sake of simplicity, in the following only this simplified model will be used. Note that figure 2 also explicitly introduces noise/interference by the superimposition of $e(n)$ in the
Figure 1: General architecture of ADC based true-RNG.

Figure 2: Single stage ADC based true-RNG.

loop. Assuming that the voltage \( v \) is meant to belong to the normalized range \([-1, 1]\), the sub-ADC computes the binary word \( D \) as

\[
d_0 = \chi_{[-1/2,0]}(\tilde{v}) \quad d_1 = \chi_{[-1/2,1]}(\tilde{v})
\]

where \( \chi(\cdot) \) denotes the characteristic function of the set in subscript and, due to noise/interference:

\[
\tilde{v}(n) = v(n) + e(n)
\]

No particular assumption is made on \( e(n) \), neither on its Probability Density Function (PDF), nor on its Power Density Spectrum (PDS). Solely, it is required to \( e(n) \) to be independent from \( v(n) \). In other terms if \( e(n) \) is a crafted interference, it must be crafted without any knowledge of \( v(n) \), which is reasonable, being that \( v(n) \) is meant to be an hidden state variable.

Thanks to the digital-to-analog conversion the \( \tilde{v}(n) \) to \( v(n + 1) \) relationship is given by a map \( M \):

\[
v(n + 1) = M(\tilde{v}(n)) =
\begin{cases}
2\tilde{v}(n) + 2 & \tilde{v}(n) \in [-1/2, -1/2] \\
2\tilde{v}(n) & \tilde{v}(n) \in [-1/2, 1/2] \\
2\tilde{v}(n) - 2 & \tilde{v}(n) \in [1/2, 1/2] \\
\text{undefined} & \text{elsewhere}
\end{cases}
\]

as also illustrated by the strong line in the plot of figure 3. It is now worth recalling what happens when \( e(n) = 0 \), namely in absence of noise. In this case \( v(n + 1) = M(\tilde{v}(n)) \). The map \( M \) is PWAM and as such the dynamics of the autonomous system in figure 2 can be coarsely modeled by a Markov chain, following the theory in [9]. Figure 4 shows the chain: the system is in state “0” when \( b = d_0 @ d_1 = 0 \) (i.e. \( x \in I_0 = ]-\infty, -1/2[ \cup ]1/2, \infty[ \)) and in state “1” when \( b = 1 \) ( \( x \in ]-1/2, 1/2[ \) ). With this it is obvious that the system generates true-random bits as the chain is identical to that of a coin toss.

3. Effects of value-limited noise

With the above considerations, the question of what happens when noise enters the game (\( e(n) \neq 0 \)) can now be tackled. To do so, it is necessary to express all quantities in statistical terms, taking advantage of their PDFs. If the PDF of \( \tilde{v} \) is indicated as \( \rho_{\tilde{v}}(x) \), the PDF of \( v \) is indicated as \( \rho_v(x) \), and the PDF of \( e(n) \) is indicated as \( \rho_e(x) \), as long as the noise samples are independent from the system state, one must have:

\[
\rho_v(x) = \text{PF}_E[\rho_{\tilde{v}}](x) = \rho_{\tilde{v}}(x) \ast \rho_e(x)
\]

where the operator \( \ast \) indicates convolution and the notation \( \text{PF}_E[\cdot] \) is used to indicate the function operator that corresponds to the addition of noise in the PDF domain.

From equation (3) one also has:

\[
\rho_v(x) = \text{PF}_M[\rho_{\tilde{v}}](x) = \int_{-\infty}^{\infty} \delta(M(\xi) - x) \rho_{\tilde{v}}(\xi) \, d\xi
\]

where \( \delta(\cdot) \) is the Dirac pulse and the operator \( \text{PF}_M \) is in fact the Perron-Frobenius operator associated to \( M \) [9].

By simultaneously solving (4) and (5), one can get the invariant PDFs of \( v \) and \( \tilde{v} \). The problem is generally known as the analysis of a \textit{constantly applied stochastic perturbation of a discrete time system}. In general terms, simultaneously solving two integral equations like (4) and (5) could be prohibitive, hence researchers have mostly focused on limit conditions, such as showing that for maps satisfying certain general properties the stationary PDF can be preserved as long as the stochastic perturbations are small [10, chapter 10]. Here, no assumption of \textit{small perturbations} shall be made. Rather, the particular nature of \( M \) will be exploited to make the simultaneous computation of (4) and (5) relatively easy. It is first worth remembering that:

**Property 1 (Map invariant density).** If \( \rho_v(x) = \delta(x) \), then
\[ \rho_v(x) = \rho_v(x) \] and
\[ \rho_v(x) = \frac{1}{2} \cdot \chi_{[0,1]}(x) \] (6)

**Proof.** The condition \( \rho_e(x) = \delta(x) \) actually corresponds to absence of noise, so the first statement is immediately proved. For what concerns the second part, note that expression (6) indeed corresponds to the invariant density of \( M \), as it can be easily seen by substituting it in (5). \( \Box \)

As a second consideration, note that:

**Property 2** (Preservation of the stationary density at the map output). Provided that \( e \in [-1/2, 1/2] \) and that \( e(n) \) remains independent from \( v(n) \), expression (6) is preserved even if \( \rho_v(x) \neq \delta(x) \).

**Proof.** This is a consequence of the fact that \( M \) satisfies the periodicity conditions expressed in [11, Theorem 1]. \( \Box \)

Interestingly, thanks to the particular observation/quantization function chosen to generate the RNG output from the chaotic system state, also the following property applies:

**Property 3** (Preservation of the statistical distribution of the coarse state). Under the same condition of the property above, if \( b = d_1 \oplus d_2 \), with \( d_1 \) and \( d_2 \) defined as in equation (1), in the presence of noise also the probability of finding the system in either the \( b = 1 \) or the \( b = 0 \) coarse state is preserved, so that \( \Pr(b = 0) = \Pr(b = 1) = 1/2 \).

**Proof.** The probability of having \( b = 1 \) is given by

\[ \Pr(b = 1) = \int_{-1/2}^{1/2} \rho_v(x) \, dx = \frac{1}{2} \int_{-1/2}^{1/2} \rho_v(x) \, dx \] (7)

Note that the inner integral is 1 for any \( x \) in \([ -1/2, 1/2 ] \), so that \( \Pr(b = 1) = \Pr(b = 0) = 1/2 \), independently from the noise PDF. \( \Box \)

Note that property 3 is not a trivial consequence of property 2. In other terms, satisfaction of the prerequisites for [11, Theorem 1] is not by itself a sufficient condition for property 3 to hold. This is because [11, Theorem 1] only regards the properties of \( \rho_v \). To conserve the probability of finding the system in a specific coarse state, also suitable conditions on the computation of \( b \) need to be satisfied, since the coarse state is obtained by quantizing \( \hat{y} \), not \( v \). For instance, if the output function was chosen to be \( \hat{b}(n) = \chi_{[0,\infty)}(x(n)) \), in absence of noise one would still have had an unbiased RNG. However, in presence of noise the probability of having \( \hat{b} = 1 \) would have become \( \frac{1}{2} \int_{-1/2}^{1/2} \rho_v(x) \, dx \), possibly different from \( 1/2 \), thus leading to bias\(^1\).

Eventually, something can be said also about the ability to model the evolution of the coarse state by a Markov chain:

\(^1\)In this case \( \Pr(b = 1) \) is certainly \( 1/2 \) only when \( \rho_v(x) \) is even.

**Property 4** (Preservation of the Markov-property for the coarse state). Under the conditions of the above properties, the possibility of modeling the coarse system behavior by a Markov chain is preserved, namely:

\[ \Pr(b(n + 1) = y \mid b(n), b(n - 1), b(n - 2), \ldots) = \Pr(b(n + 1) = y \mid b(n)) \] (8)

**Proof.** First of all, it is worth considering that \( \Pr(b(n + 1) = y \mid \ldots) \) is given by the knowledge of \( \rho_{\hat{b}(n+1)}(\cdot) \) i.e. by the knowledge of function \( \rho_{\hat{b}} \) at time \( n + 1 \), when suitably conditioned,\(^2\) as in

\[ \Pr(b(n + 1) = y \mid \ldots) = \int_{b} \rho_{\hat{b}(n+1)\mid b(n)}(\cdot) \, dx \] (9)

Hence, to prove (8), it is sufficient to prove that

\[ \rho_{\hat{b}(n+1) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) = \rho_{\hat{b}(n+1) \mid b(n)}(\cdot) \] (10)

It shall first be shown that for (10) to hold it is sufficient to have

\[ \rho_{\hat{b}(n+1) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) = \rho_{\hat{b}(n+1) \mid b(n)}(\cdot) \] (11)

By usage of the PF\(_E\) and PF\(_M\) operators one has:

\[ \rho_{\hat{b}(n+1) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) = \rho_{\hat{b}(n+1) \mid b(n)}(\cdot) \] (12)

In other terms, \( \rho_{\hat{b}(n+1) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) \) is fully determined by \( \rho_{\hat{b}(n) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) \). Focus shall thus be on the latter, which contains a conditioned probability density function where the conditioning variable is defined on a discrete set. To deal with this form, a special case of the Bayes expression shall be used:

\[ \rho_{\hat{b}(n) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) = \frac{\rho_{\hat{b}(n)}(\cdot) \Pr(b(a = x) \mid b(n))}{\Pr(b)} \] (13)

where \( a \) is a variable defined on \( \mathbb{R} \) and \( b \) is a variable defined in a discrete set. By applying (13) one gets:

\[ \rho_{\hat{b}(n) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) = \frac{\rho_{\hat{b}(n)}(\cdot) \Pr(b(a = x) \mid b(n))}{\Pr(b)} \] (14)

In other terms, by equations (12) and (14), the PDF \( \rho_{\hat{b}(n+1) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots}(\cdot) \) is recursively defined from knowledge of \( \rho_{\hat{b}(n)}(\cdot) \mid b(n) \mid b(n-1) \mid b(n-2) \ldots \mid b(n-3) \ldots \mid b(n-2) \ldots \mid b(n-3) \ldots \mid b(n-2) \ldots \mid b(n-3) \ldots \mid b(n-2) \ldots \mid b(n-3) \ldots \) and of \( b(n) \). If the first is not dependent on \( b(n - 1) \), then the second, having the same structure, cannot be dependent on \( b(n - 2) \). Hence, the first cannot be dependent on \( b(n - 2) \) too. But if this

\(^2\)The standard notation for the PDFs of variable \( z \) conditioned by \( y \) is \( \rho_{z \mid y}(x) \). Here a different notation is used, i.e. \( \rho_{\hat{b} \mid y}(x) \). This is to emphasize that the conditioned PDF is a function (dependent on) of a single argument \( x \). This notation is particularly useful to maintain a consistent style when applying functional operators on the conditioned PDF.
holds, the second cannot be dependent on $b(n - 3)$ and so on...

Having established the above, (11) can now be proved. Suppose that it is known that $b(n - 1) = z_0 - 1$ (and nothing else). Then, by (13), one has

$$
\rho_{b(n-1) ; b(n-2)} (v) = \frac{\rho_e (x) \cdot \chi_{I_{n-1}} (x)}{Pr (b = z_{n-1})} = 2 \int_{\chi_{I_{n-1}}}^{\chi_{I_{n-1}}} \rho_e (x) \cdot \chi_{I_{n-1}} (x) \cdot \chi_{I_{n-1}} (x) \, dx
$$

Eventually, by transforming the above using PF$_M$, one has

$$
\rho_{b(n-1) ; b(n-2)} (v) = \frac{1}{2} \left[ \int_{\chi_{I_{n-1}}}^{\chi_{I_{n-1}}} \rho_e (x) \cdot \chi_{I_{n-1}} (x) \cdot \chi_{I_{n-1}} (x) \, dx + \int_{\chi_{I_{n-1}}}^{\chi_{I_{n-1}}} \rho_e (x) \cdot \chi_{I_{n-1}} (x) \cdot \chi_{I_{n-1}} (x) \, dx \right]
$$

$$
\int_{\chi_{I_{n-1}}}^{\chi_{I_{n-1}}} \rho_e (x) \cdot \chi_{I_{n-1}} (x) \cdot \chi_{I_{n-1}} (x) \, dx \, dx
$$

For $z_{n-1} = 1$, only the central integral is non-null, so the expression evaluates to $1/2$. Similarly, for $z_{n-1} = 0$, only the first and last integrals are non-null and they globally evaluate to $1/2$. Hence, the above reduces to (6) independently from the particular $b(n - 1)$. In other terms, in one iteration memory of $b(n - 1)$ is completely lost. With this, also $\rho_{b(n-1) ; b(n-2)} (v)$ must be independent from $b(n - 1)$ and, by (12) equation (11) is proved. □

4. Application to AWGN and interference

Properties 3 and 4 basically state that in presence of value-limited noise an ADC-derived true-RNG built from ideal parts (namely respecting equation (3) strictly) remains a perfect, true-RNG. It is thus interesting to apply this result to real-world situation.

The first case that shall be considered is robustness to Additive White Gaussian Noise (AWGN) which exists in every electronic circuit due to Johnson (thermal) effects. Strictly speaking, AWGN does not satisfy the preconditions of properties 2 to 4, since $\rho_{\text{AWGN}} (x) > 0, \forall x \in \mathbb{R}$, independently from the standard deviation $\sigma_{\text{AWGN}}$. However, it is intuitive that for low $\sigma_{\text{AWGN}}$ the probability of a noise sample invalidating the above analysis is actually very low. Simple computations show that even for a system with a mediocre Signal to Noise Ratio (SNR) of 20 dB, the probability of breaking the system behavior is already much less than 10$^{-19}$ (i.e. absolutely negligible). To make the system somehow robust to really large AWGN, one needs to extend the map $M$ as shown by the dotted line in fig. 3. With this, property 2 remains satisfied, however the remaining properties do not and one does not have an ideal true-RNG anyway.

A second case to consider is robustness to external interference (either unintentional or intentional). Note that in properties 2 to 4 there is no mention that $e(n)$ must be stationary. Indeed, all the above theory continues to hold even if $\rho_e (x)$ is time-variant. The only condition on $e(n)$ is that it is independent from $v(n)$.

4.1. Example/validation

As an example/validatio a simulator has been programmed, considering the architecture of figure 2, with a large sinusoidal interference with amplitude 0.2 peak-to-peak. Quality of the output has been evaluated by estimating the average entropy per bit of 10$^6$ output bits, by looking at 16-bit packets. This procedure resulted in 1 bit of entropy per bit up to 5 decimal digits of precision, i.e. in no influence on the quality of the output stream from the interfering signal.

References


3For instance, $e(n) = A \sin(2\pi f n)$ can be modeled through a time-variant $\rho_{v}(x) = \delta(x - A \sin(2\pi f n))$. 

390
Compensated True Random Number Generator Based On a Double-Scroll Attractor

Salih Ergün† and Serdar Özoğuz‡

†TÜBİTAK-National Research Institute of Electronics and Cryptology
PO Box 74, 41470, Gebze, Kocaeli, Turkey
‡İstanbul Technical University, Faculty of Electrical-Electronics Engineering
34390, İstanbul, Turkey
Email: salih@uekae.tubitak.gov.tr, serdar@ebu.itu.edu.tr

Abstract—A novel random number generator (RNG) based on a double-scroll attractor is presented. Simulation and experimental results, showing that the generated binary sequence pass standard randomness statistical tests without post processing and verifying the feasibility of the circuit, are given. In the proposed RNG, offset and frequency compensation loops are added to maximize the statistical quality of the output sequence and to be robust against parameter variations and attacks. The proposed RNG can be realized in integrated circuit.

1. Introduction

Nowadays, because of the increasing demand of electronic official & financial transactions and digital signature applications, the need for information secrecy has raised. In this manner, random number generators (RNGs) which have been used for only military cryptographic applications in the past got expanding usage for a typical digital communication equipment.

There are few RNG designs reported in the literature, in spite of the fact that, the use of discrete-time chaotic maps [1, 2] in the realization of RNG is well-known for some time, it was only recently shown that continuous-time chaotic oscillators [3, 4] can be used to realize truly random number generators (TRNGs) also. Following up in this direction, we investigated the usefulness of the proposed RNG design to generate random binary data from continuous-time chaotic oscillators.

Initially, we have obtained random data by periodically sampling one of the state of the chaotic system and numerically verified that the bit streams generated from the proposed RNG pass the four basic randomness tests of FIPS-140-2 test suite [5]. External interference is a major concern in RNG design since interfered and random signals have comparable levels. To solve this problem and to be robust against parameter variations and attacks aimed to force throughput, we have proposed offset and frequency compensation loops that increase the statistical quality of the generated bit sequences. Finally we have experimentally verified that the binary data obtained from the proposed circuit pass the tests of NIST full random number test suite [6] without post processing.

2. Double-Scroll Attractor and Random Bit Generation

The double-scroll attractor which is used as the core of the RNG is obtained from a simple model given in [7], which is expressed by the Equation 1.

\[
\begin{align*}
\dot{x} &= y \\
\dot{y} &= z \\
\dot{z} &= -ax - ay - az + \text{sgn}(x)
\end{align*}
\] (1)

where \(\text{sgn}(.)\) is the signum function. The equations in 1 generate chaos for different set of parameters. For example, the chaotic attractor shown in Fig.1a is obtained from the numerical analysis of the system with \(a = 0.666\) using a 4th-order Runge-Kutta algorithm with an adaptive step size.

![Figure 1: (a) Numerical analysis; (b) Experimental results of the chaotic oscillator.](image)

In order to obtain random binary data from a continuous-time chaotic system, we have presented an interesting technique, which relies on generating a non-invertible binary data from the waveform of the given chaotic system. It should be noted that non-invertibility is a key feature for generating pseudo random numbers [8].
To obtain binary random bits from the chaotic attractor, we used the samples of the state $x$ of the system in Equation 1, obtained at the rising edges of an external periodical pulse signal, that is at times $t$ satisfying $\text{wtmod}2\pi = 0$ where $w$ is the frequency of the pulse signal. Note that, although 3-dimensional trajectories in the $x-y-z$ plane is invertible, one may obtain a non-invertible section by considering only the values corresponding to one of the states, say $x$.

We initially examined the distribution of periodically sampled $x$ values to determine appropriate sections where the distribution looks like random signal. Although, we could not find sections of which $x$ has a single normal or $\chi^2$-distribution for any $a$ given in Equation 1, we determined various sections where the distribution of $x$ has at least two regions. This suggests us to generate random binary data from regions of $x$ values for regional thresholds. Following this direction, we have generated the binary data $S_{\text{top}}(i)$ and $S_{\text{bottom}}(i)$ for $a = 0.666$ from the 1-dimensional section according to the Equation 2:

$$
S_{\text{top}}(i) = \text{sgn}(x_i - q_{\text{top}}) \quad \text{when } x_i \geq q_{\text{middle}}
$$

$$
S_{\text{bottom}}(i) = \text{sgn}(x_i - q_{\text{bottom}}) \quad \text{when } x_i < q_{\text{middle}}
$$

where $x_i$'s are the values of $x$ at the 1-dimensional section obtained for $\frac{1}{3}\text{wtmod}2\pi = 0$ ($v = 1/2$). $q_{\text{top}}$ and $q_{\text{bottom}}$ are the thresholds for top and bottom distributions, respectively and $q_{\text{middle}}$ is the boundary between the distributions. To be able to choose the thresholds appropriately, we examined top and bottom distributions and then, $q_{\text{top}}$ and $q_{\text{bottom}}$ were determined as the medians of the top and bottom distributions which were 0.9656158849 and $-0.9640518966$, respectively when $q_{\text{middle}}$ was determined as 0.

Generation of the binary sequence thus obtained does not so much dependent on $q_{\text{middle}}$ value, because for this boundary value, distribution density of $x$ is minimum. However, distribution density of $x$ for threshold values ($q_{\text{top}}, q_{\text{bottom}}$) is maximum so the binary sequence obtained may be biased. In order to remove the unknown bias in this sequence, we have used Von Neumann’s de-skewing technique [10] in the following manner:

$$
S_{\text{top}}(i) \text{ XOR } S_{\text{bottom}}(i)
$$

where $S_{\text{top}}(i)$ and $S_{\text{bottom}}(i)$ are the output bit stream of the comparators were sampled and stored in binary format at the

We calculated as 0.00018 and it is determined that the generated binary sequences are independent. This was, in fact, expected as the chaotic systems are characterized by having a positive Lyapunov exponent [12], and the auto-correlation of the chaotic time-series vanish abruptly. According to this result, we have generated the new binary data $S_{\text{cor}}(i) = S_{\text{top}}(i) \text{ XOR } S_{\text{bottom}}(i)$.

The mean value $\psi$ of the binary sequence $S_{\text{cor}}$ thus obtained, can be calculated by: $\psi = \frac{1}{2} - 2(\mu - \frac{1}{2})(\nu - \frac{1}{2})$, where the mean value of $S_{\text{top}} = \mu$ and the mean value of $S_{\text{bottom}} = \nu$. Thus if $\mu$ and $\nu$ are close to $\frac{1}{2}$ then $\psi$ is very close to $\frac{1}{2}$. As a result, we have numerically verified that the bit sequence $S_{\text{cor}}$, which was obtained for the given appropriate threshold values, passed the tests of FIPS-140-2 test suite without Von Neumann processing. We called random number generation according to above procedure, as Regional – RNG.

3. Hardware Realization of RNG and Test Results

Due to the lack of access to a suitable fabrication facility, we have chosen to construct the proposed RNG circuit using discrete components in order to show the feasibility of the circuit.

The double-scroll attractor circuit was realized by using the model given in [7]. AD844 was used as a high speed operational amplifier and LM211 voltage comparator was used to realize required nonlinearity. The passive component values given in [7] were taken as: $R = aR_S = 10k\Omega$, tuning resistor $R_S = 15k\Omega$ for $a = 0.666...$, and $C = 2.2n\mu F$. Therefore the center operation frequency of the chaotic oscillator: $f = \frac{1}{\pi \tau}$ corresponding to time constant $\tau$ where $\tau = R_C$, was adjusted to a low frequency value as 7.234 kHz on purpose to provide the circuit not to be affected by parasitic capacitances. The circuit was biased with a $\pm 5V$ power supply and the observed attractor is shown in Fig.1b.

In Regional RNG, the voltage $v_1$, which corresponds to the variable $x$, was converted into binary sequences according to the procedure explained in Section 2. In order to implement this procedure, the circuit shown in Fig. 2 was used. In this circuit, the comparators were implemented from LM211 chips and the voltage levels $V_{\text{top}}$, $V_{\text{middle}}$ and $V_{\text{bottom}}$ were used to realize the thresholds in Equation 2, respectively. $V_{\text{top}}$ and $V_{\text{bottom}}$ were generated by two 12-bit voltage-mode digital to analog converters (DACs). Each DAC can be adjusted in 0.58559375 mV steps where the voltage reference of the DAC is 2.4V.

An FPGA based hardware, which has a PCI interface was designed to upload the binary data to the computer. To obtain $x$ values in the section defined as $\text{wtmod}2\pi = 0$, output bit stream of the comparators were sampled and stored in binary format at the
Figure 2: Regional RNG.

rising edge of the external periodical square-wave generator \( v_p(t) \). Offset compensation for \( V_{top} \) and \( V_{bottom} \) thresholds, frequency compensation and exclusive-or operation were implemented inside the FPGA. After offset & frequency compensation and exclusive-or operation, the candidate random numbers were uploaded to the computer through the PCI interface. Maximum data storage rate of our FPGA based hardware is 62 Mbps.

According to the procedure explained in Section 2, we examined the distribution of \( v_1 \). As a result, similarly to numerical bit generation, distribution of \( v_1 \) obtained at the rising edge of \( v_p(t) \) has two separate regions. To be able to determine initial values of the thresholds appropriately top and bottom distributions were examined. Then, initial values of \( V_{top} \) and \( V_{bottom} \) were determined as the medians of the top and bottom distributions which were \( 470mV \) and \( -470mV \), respectively while \( V_{middle} \) was determined as \( 0mV \). Sampling frequency of \( v_1 \), was determined by dividing the frequency of \( v_p(t) \) into pre-scaler value inside the FPGA.

Figure 3: Frequency spectrum of \( v_1 \).

For determining the initial value of the pre-scaler appropriately, the frequency spectrum of \( v_1 \) given in Fig. 3 was observed. As shown in Fig. 3 chaotic signal \( v_1 \) has noise-like power spectrum. The center frequency of the chaotic oscillator is indicated by the solid marker set at \( 7.234 \text{ KHz} \). Up to the dashed marker set at \( 1.55 \text{ KHz} \), the region in which the power spectrum is flat, chaotic signal \( v_1 \) contains all frequencies in equal amounts and power spectral density is at its maximum. Hence, without loss of generality, \( v_1(t) \) and \( v_1(t + t_0) \) can be considered as uncorrelated for all \( t_0 \neq 0 \) and \( v_1 \) can be sampled up to \( 1.55 \text{ KHz} \) as a random signal source. Finally, initial value of the pre-scaler was determined as \( 3 \), while the frequency of \( v_p(t) \) was \( 4.65 \text{ KHz} \).

Offset compensations of \( V_{top} \) and \( V_{bottom} \) thresholds were realized by implementing monobit test of FIPS-140-2 test suite [5] for \( S_{top} \) and \( S_{bottom} \) binary sequences. For each sequence, bit streams of length 20000 Bits were acquired, if the number of 0’s > 10275 then corresponding threshold was decreased and if the number of 0’s < 9725 then corresponding threshold was increased.

Figure 4: The effect of offset compensation for \( V_{top} \).

Frequency compensation loop was realized by implementing runs test of FIPS-140-2 test suite for \( S_{xor} \) binary sequence. If \( 3 \) \( S_{xor} \) bit streams of length 20000 Bits which were acquired in sequence failed in runs test, which indicated over sampling of \( v_1 \), then sampling frequency of \( v_1 \) was scaled down by increasing the pre-scaler value. Pre-scaler value, initial value of which was determined as \( 3 \), became stable at \( 4 \). If necessary, sampling frequency can be scaled up externally through the PCI interface. The effect of offset compensation for \( V_{top} \), which is similar to one for \( V_{bottom} \), is shown in Fig. 4, in spite of the fact that, initial value of the threshold was not adjusted appropriately, mean value of the bit stream of length 20000 Bits reached and became stable at \( 4 \) thanks to compensation.

After pre-scaler and threshold values became stable, bit stream of length 223 MBits was acquired and subjected to full NIST test suite. As a result, we have experimentally verified that, bit sequence \( S_{xor} \), passed the tests of full NIST random number test suite without Von Neumann processing. Results for the uniformity of P-values and the proportion of passing sequences of the Regional RNG circuit are given in Table 1 where P-value \( (\leq P-value \leq 1) \) is a real number estimating the probability that a perfect RNG would have produced a sequence less random than the given sequence. It is reported that, for a sample size of \( 223 \times 1MBits \), the minimum pass rate for each statistical test with the exception of the random excursion (variant) test is approximately 0.970011.

When the center frequency of the chaotic oscillator was \( 7.234 \text{ KHz} \) throughput data rate of \( S_{xor} \) effectively becomes \( 581\text{bps} \). (Pre-scaler value: 4) because of dividing \( v_1 \) into two regions according to distribution. Throughput data rate of \( S_{xor} \) can be generalized as \( f_{xor} = \frac{28.5K\Omega}{\tau_{new}} = \frac{1}{\tau_{new}} \), where \( \tau_{new} = R_{new}C_{new} \). In [7], a chip realization of the double-scroll system with \( R_{new} = 28.5K\Omega \) and \( C_{new} = 15pF \) has been presented, which leads to a center frequency of operation at \( f = \frac{1}{2\pi\tau_{new}} = 500\text{KHz} \). Consider-
ing that the circuit in [7] was realized on a relatively slow 1.2u CMOS process, we can deduce that the circuit can easily be integrated on today’s process at a couple of 10MHz and can generate throughput closer to Mbps. However, it should be noted that chaotic circuits operating at much higher frequencies are reported in literature.

<table>
<thead>
<tr>
<th>STATISTICAL TESTS</th>
<th>$S_{stat}$ Bit Sequence</th>
<th>$P$-Value</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0.084879</td>
<td>0.9955</td>
<td></td>
</tr>
<tr>
<td>Block Frequency</td>
<td>0.320612</td>
<td>0.9824</td>
<td></td>
</tr>
<tr>
<td>Cumulative Sums</td>
<td>0.136566</td>
<td>0.9955</td>
<td></td>
</tr>
<tr>
<td>Run</td>
<td>0.392496</td>
<td>0.9776</td>
<td></td>
</tr>
<tr>
<td>Longest Run</td>
<td>0.298151</td>
<td>0.9865</td>
<td></td>
</tr>
<tr>
<td>FFT</td>
<td>0.231847</td>
<td>0.9865</td>
<td></td>
</tr>
<tr>
<td>Nonperiodic Templates</td>
<td>0.710974</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Overlapping Templates</td>
<td>0.053938</td>
<td>0.9776</td>
<td></td>
</tr>
<tr>
<td>Universal</td>
<td>0.941444</td>
<td>0.9955</td>
<td></td>
</tr>
<tr>
<td>Apen</td>
<td>0.449956</td>
<td>0.9824</td>
<td></td>
</tr>
<tr>
<td>Random Excursions</td>
<td>0.725540</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Random Excursions Variant</td>
<td>0.901761</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>Serial</td>
<td>0.744459</td>
<td>0.9955</td>
<td></td>
</tr>
<tr>
<td>Linear Complexity</td>
<td>0.797289</td>
<td>0.9865</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: NIST test suite results of the Regional RNG.

While comparing our Regional RNG design with the previous one given in [3], we have experimentally verified that, for the same chaotic oscillator the through-put data rate of the RNG method given in [3] was 385bps. Furthermore bit sequence obtained from the RNG method given in [3] can pass the full test suite of Diehard with only Von Neumann processing.

As a result, the proposed design is an enhanced architecture where offset and frequency compensation loops are added to maximize the statistical quality of the output sequence and to be robust against external interference, parameter variations and attacks aimed to force throughput.

4. Conclusions

A novel TRNG based on a continuous-time chaotic oscillator was presented. Numerical and experimental results presented in this paper not only verify the feasibility of the proposed circuit, but also encourage its use as the core of a high-performance IC RNG as well. In comparison with RNGs based on discrete-time chaotic maps, which are advantageous in the sense that true random behavior can be mathematically proven thanks to an analytical model that has been developed, it is seen that RNGs based on continuous-time chaotic oscillators can offer much higher and constant data rates without post-processing. In conclusion, we can deduce that the use of continuous-time chaos is very promising in generating random numbers with very high throughput.

References

A low cost chaos-based random number generator realized in 8-bit precision
environment

K.W. Tang and Wallace Tang

Department of Electronic Engineering
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong, China
Email: khtang, kstang@ee.cityu.edu.hk

Abstract—Random number generators (RNGs) are widely used in different applications, but are difficult to be realized in low precision and resource constrained system, such as the 8-bit micro-controller systems. This paper provides a feasible solution for such a controller system, which is commonly used in consumer and industrial markets. By sampling a modified Chua’s circuit and post-processing with a high-dimensional chaotic map, a RNG is designed. The randomness of the generated bit sequence is also testified and confirmed by the up-to-date statistical suite.

1. Introduction

Low cost micro-controllers and micro-processors are dominated in industrial and consumer markets. However, their low precision and limited resources have largely hindered the implementation of random function, causing some trade-offs or limitation in applications. An example is the “random shuffle” function in the music devices. Although songs are supposed to be selected randomly, in fact, the sequence is not random. Moreover, the quality of the RNG will also affect the performance of some algorithms, for example evolutionary algorithms, encryption algorithms and so on.

Implementing RNG in a low-precision controller is not an easy task. Severe problems such as short cycle length, non-ideal distribution and correlation functions have been noticed [3]. Although the use of chaos in random number generation has been proposed by a number of researchers [7, 8, 9], a practical implementation under 8-bit precision has never been investigated.

In this paper, a low cost chaos-based RNG is to be designed. The organization of the paper is as follows. Firstly, the design of the proposed RNG is explained in detail in Sect. 2. Its performance study is given in Sect. 3. Finally, conclusion remarks are drawn in Sect. 4.

2. Design of Random Number Generator

The use of 8-bit micro-controllers and micro-processors are common in industry and consumer markets due to their low costs. However, it is a great challenge to have RNG realized under such a low precision environment, although a good RNG is always demanded for lots of applications. In this paper, a chaos-based RNG is proposed for the 8-bit precision environment. The block diagram is depicted in Fig. 1, and it consists of two major parts. Firstly, bit stream is generated by sampling a chaotic circuit, and then random bit sequence can be obtained by a simple post-processing method.

![Figure 1: Block diagram of chaos-based RNG design](image)

2.1. Modified Chua’s Circuit

In order to generate the bit stream for the RNG, a very simple chaotic circuit proposed in [2] is adopted. It is a modified Chua’s circuit, which can be expressed as follows:

\[ C_1 \frac{dV_1}{dt} = \frac{1}{R} (V_2 - V_1) - f(V_1) \]
\[ C_2 \frac{dV_2}{dt} = \frac{1}{R} (V_1 - V_2) + I_3 \]
\[ L \frac{dI_3}{dt} = -V_2 - I_3 R_0 \]  

(1)

where \( f(V) = G_a V - \frac{G_{max}}{V_{max}} V^3 \) is a cubic-like voltage-current characteristic [10] of the nonlinear resistor with slopes \( G_a \) and \( \pm V_{max} \) are the outer zeros.

The overall circuit is depicted in Fig. 2 where the shaded part is used to realize the \( f(V) \). The double-scroll chaotic attractor is then obtained as in Fig. 3 with the cubic v-i function \( f(V) \) shown in Fig. 4.

The use of this circuit in our RNG design has the following advantages:

1. The circuit is extremely simple and only consists of a few passive components and two pairs of dual complementary MOS transistors.
2. A chaotic attractor is observed, and hence non-periodic bit sequence can be obtained after data sampling.

2.2. Data Postprocessing

Referring to Fig. 1, one of the system states (in our design, $V_2$) of the modified Chua’s circuit is sampled and used as the bit stream. However, even though the sampled data is non-periodic, its randomness may not be good enough for RNG applications (that is confirmed in our test, which will be discussed further in Sect. 3). Therefore, data post-processing is necessary.

In our approach, a chaotic map, called high dimensional Cat map, is constructed as the post-processing function because of its nice simple structure. This map [1] is found to possess a very good mixing property, even under finite precision. Moreover, it can be easily implemented in a 8-bit precision environment.

Firstly, let’s define an extended Cat map as follows:

$$z(k + 1) = A_{12}z(k)$$  \hspace{1cm} (2)

where $z(k) = [z_1 \ z_2 \ \cdots \ z_n]^T$, and $A_{12}$ is an $n \times n$ matrix expressed as

$$A_{12} = \begin{bmatrix} 1 & \alpha & 0 & \cdots & 0 \\ \beta & 1 + \alpha\beta & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$  \hspace{1cm} (3)

Similar to the conventional two-dimensional Cat map, the 1st and 2nd dimensions of $z$ are mixed, according to the mapping matrix $A_{12}$ specified by $\alpha$ and $\beta$.

The high-dimensional Cat map is then formed as:

$$z(k + 1) = Az(k)$$  \hspace{1cm} (4)

where $A$ is an $n \times n$ matrix, constructed by

$$A = A_{12}A_{13} \cdots A_{1n}A_{23} \cdots A_{2n} \cdots A_{(n-1)n}$$  \hspace{1cm} (5)

with each $A_{ij}$ mixing the $i$-th and $j$-th dimensions as in the extended Cat map given in Eq. (2) and (3). Therefore, with the mapping matrix, $A$, all the dimensions of $z$ will be mixed completely.

In our experiment, $n = 5$, and it should be emphasized that $A$ is predefined and hence is just a constant matrix. The bit stream obtained by sampling the chaotic circuit is fed into this high dimensional Cat map, Eq. (4), in groups of n-byte (i.e. each byte obtained from a data sample), and then processed by the map.

2.3. Hardware Implementation

The block diagram of the proposed chaos-based RNG system and a prototype are shown in Fig. 5 (a) and (b), respectively. The modified Chua’s circuit in Eq. (1) is realized by the passive components (L, C, and R), together with
CD4007 for the realization of $f(V)$. The analog signal of
the state $V_2$ is sampled by an analog-to-digital converter
(MAX118) with a sampling rate of 50KHz and the samples
are stored in a SRAM. The post-processing function or
the high-dimensional Cat map in Eq. (4) is programmed in
the 8-bit controller $\mu$C8051, occupying 33 bytes program
memory and 35 bytes internal RAM. The size of overall
program is 259 bytes. The output is used as the random bit
sequence.

3. Performance Analysis

In order to verify the randomness of the output bit se-
quence from the $\mu$C8051, the sequence is streamed to the
computer via the COM-port by RS232 for further analysis.

3.1. Correlation Functions

Firstly, the auto-correlation and cross-correlation of
the sequences are studied and exhibited in Figs. 6 and
7, respectively. It clearly shows that the use of high-
dimensional Cat map can improve their performance,
achieving good auto-correlation and cross-correlation
properties.

3.2. Bit Balance Probability

Let $p$ and $q$ be the frequencies of “0” and “1”, respec-
tively, then the quantity $\frac{p-q}{N}$ implies the balance of these
probabilities, where $N$ is the number of bits. Fig. 8 de-
picts the balance of probabilities and a very good result
is obtained. The result for the bit sequence without post-
processing is also depicted for comparison.

3.3. Statistical Test Suite

The proposed scheme is also evaluated by the NIST test
suite [4], which contains fifteen statistical tests, covering
a wide area of FIPS 140-2 [5]. The tests can be classified
into three major random natures:

1. Random Walk Nature: frequency test (FT), block fre-
cuency test (BFT), cumulative sums test (CST), ran-
dom excursions test (RET), random excursions variant
test (REVT).
2. Pattern Checking: runs test (RT), longest run of
ones test (LROT), nonoverlapping template match-
ing test (NTMT), overlapping template matching test
(OTMT), Maurer’s universal statistical test (MUST),
approximate entropy test (AET), serial test (ST).
3. Complexity and Compression: Marsaglia’s rank test
(MRT), spectral test (SPT), linear complexity test
(LCT).
Details of the tests can be referred to [6].

The results are tabulated in Table 1 and it is found that all the tests are passed. The results using sampled data from the modified Chua’s circuit without post-processing are also listed for comparison.

<table>
<thead>
<tr>
<th>Tests</th>
<th>Passing rate without post-processing</th>
<th>Passing rate with post-processing</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>0.00%</td>
<td>97.67%</td>
</tr>
<tr>
<td>BFT</td>
<td>0.00%</td>
<td>100.00%</td>
</tr>
<tr>
<td>CST(a)</td>
<td>0.00%</td>
<td>98.00%</td>
</tr>
<tr>
<td>CST(b)</td>
<td>0.00%</td>
<td>97.67%</td>
</tr>
<tr>
<td>RET</td>
<td>0.00%</td>
<td>99.17%</td>
</tr>
<tr>
<td>REVET</td>
<td>0.00%</td>
<td>99.40%</td>
</tr>
<tr>
<td>RT</td>
<td>0.00%</td>
<td>98.67%</td>
</tr>
<tr>
<td>LROT</td>
<td>0.00%</td>
<td>99.33%</td>
</tr>
<tr>
<td>NTMT</td>
<td>12.70%</td>
<td>98.99%</td>
</tr>
<tr>
<td>OTMT</td>
<td>0.00%</td>
<td>98.33%</td>
</tr>
<tr>
<td>MUST</td>
<td>0.00%</td>
<td>98.33%</td>
</tr>
<tr>
<td>AET</td>
<td>0.00%</td>
<td>99.33%</td>
</tr>
<tr>
<td>ST</td>
<td>0.00%</td>
<td>99.00%</td>
</tr>
<tr>
<td>MRT</td>
<td>94.00%</td>
<td>99.33%</td>
</tr>
<tr>
<td>SPT</td>
<td>1.00%</td>
<td>98.67%</td>
</tr>
<tr>
<td>LCT</td>
<td>99.33%</td>
<td>98.33%</td>
</tr>
</tbody>
</table>

1 The length of the bit sequence under testing is 1,000,000 bits, and 300 sequences are used. Based on the standard significance levels [6], the minimum acceptable percentage in passing a test is 97.28%.

2 CST(a) and CST(b) stand for the cumulative forward sum and cumulative backward sum, respectively.

4. Conclusions

A novel chaos-based RNG is designed under a 8-bit precision environment. The core consists of a simple modified Chua’s circuit. Bit stream is obtained by sampling a state of the circuit, and then further mixed by a high-dimensional Cat map, which has been proved to be an effective function for improving the statistical quality of the bit sequence. The overall design has been implemented in a low-cost 8-bit controller, and the generated bit sequences are confirmed to pass all the tests in the up-to-date NIST test suite even under such a low precision environment.

Acknowledgments

The work described in this paper was fully supported by a grant from the Research Grants Council of Hong Kong Special Administrative Region, China, CityU 120205.

References

Computational complexity analysis of a Petri net identification procedure

Maria Paola Cabasino, Alessandro Giua, Carla Seatzu

Department of Electrical and Electronic Engineering, University of Cagliari
Piazza d'Armi, Cagliari 09123, Italy
Email: (cabasino,giua,seatzu)@diee.unica.it

Abstract—In previous papers we presented an approach to identify a Petri net system, given its language, based on the solution of an integer programming problem.

In this paper we analyze the complexity of such an approach in terms of computational time required to get an admissible solution, that may also be optimal according to a given performance criterion.

1. Introduction

This paper is based on our previous results in [1, 2]. In particular, in [1] we proposed a linear algebraic characterization of the Petri net systems that are able to generate a given finite set of strings, assuming that the number of places is known a priori. We also shown how this characterization can be used to synthesize a bounded net system whose language is given in terms of its reachability graph, namely imposing an appropriate set of finite length strings plus the set of minimal T-invariants.

In [2] we shown how the assumption that the number of places is given can be relaxed, and the approach can be generalized to the case in which the number of places is only known to be less or equal to a given value. In both cases, if a performance index is considered, the above linear algebraic characterization can be used to formulate a linear integer programming problem (IPP), that can be solved using ILOG CPLEX.

The goal of this paper is that of analyzing the computational complexity of the above approach. In particular, we want to investigate how the computational time depends on the cardinality of the set of finite length strings that describe the language, and on the chosen performance index.

To this aim we consider the language generated by a particular Petri net system that models a sender-receiver process. Different cases are examined with different number of places and transitions, and thus different languages generated.

The numerical simulations we carried out enabled us to conclude that the computational time become prohibitive for languages that are described by a large number of finite length strings, if we want to determine a solution that is optimal with respect to a given performance index. On the contrary, computational times are negligible if we limit to consider any admissible net system, e.g., the first admissible solution computed by CPLEX when solving the optimization problem. We believe that this is not a drawback of our procedure because in effect, when solving identification problems like this, the main requirement is that of determining an admissible solution, not necessarily an optimal one.

2. Background of Petri nets

In this section we recall the formalism used in the paper. For more details on Petri nets see [3].

A Place/Transition net (P/T net) is a structure $N = (P, T, \text{Pre}, \text{Post})$, where $P$ is a set of $m$ places; $T$ is a set of $n$ transitions; $\text{Pre} : P \times T \to \mathbb{N}$ and $\text{Post} : P \times T \to \mathbb{N}$ are the pre- and post-incidence functions that specify the arcs; $C = \text{Post} − \text{Pre}$ is the incidence matrix.

A marking is a vector $M : P \to \mathbb{N}$ that assigns to each place of a P/T net a non-negative integer number of tokens, represented by black dots. We denote $M(p)$ the marking of place $p$. A P/T system or net system $\langle N, M_0 \rangle$ is a net $N$ with an initial marking $M_0$.

A transition $t$ is enabled at $M$ iff $M \geq \text{Pre}(\cdot, t)$ and may fire yielding the marking $M' = M + C(\cdot, t)$. We write $M(\sigma)$ to denote that the sequence of transitions $\sigma$ is enabled at $M$, and we write $M(\sigma) M'$ to denote that the firing of $\sigma$ yields $M'$. Note that in this paper we always assume that two or more transitions cannot simultaneously fire (non-concurrency hypothesis).

A marking $M$ is reachable in $\langle N, M_0 \rangle$ iff there exists a firing sequence $\sigma$ such that $M_0(\sigma) M$. The set of all markings reachable from $M_0$ defines the reachability set of $\langle N, M_0 \rangle$ and is denoted $R(N, M_0)$.

Given a Petri net system $\langle N, M_0 \rangle$ we define its language as the set of its firing sequences $L(N, M_0) = \{ \sigma \in T^* \mid M_0(\sigma) \}$.

3. The identification procedure

In [1] we considered the following problem.

Problem 3.1 Assume we are given a set of places $P = \{p_1, \ldots, p_m\}$ and a set of transitions $T = \{t_1, \ldots, t_n\}$. Let $L \subset T^*$ be a finite prefix-closed lan-
guage\(^1\) over \(T\), and \(k = \max_{\sigma \in \mathcal{L}} |\sigma|\) be the length of the longest string in \(\mathcal{L}\).

We want to identify the structure of a net \(N = (P, T, \text{Pre}, \text{Post})\) and an initial marking \(M_0\) such that \(L_k(N, M_0) = \mathcal{L}\).

The unknowns we want to determine are the elements of the two matrices \(\text{Pre} = \{e_{i,j}\} \in \mathbb{N}^{m \times n}\) and \(\text{Post} = \{o_{i,j}\} \in \mathbb{N}^{m \times n}\) and the elements of the vector \(M_0 = [m_{0,1} \ m_{0,2} \cdots \ m_{0,m}]^T \in \mathbb{N}^m\).

In [1] we proved that a solution to the above identification problem can be computed thanks to the following theorem.

**Theorem 3.2** [1] A solution to the identification problem (3.1) satisfies the following set of linear algebraic constraints

\[
\mathcal{G}(m, T, \mathcal{L}) \triangleq \left\{ 
\begin{array}{ll}
M_0 + \text{Post} \cdot \text{Pre} \cdot (\delta + \bar{\varepsilon}) \geq 0 & \forall (\sigma, t_j) \in \mathcal{E} \\
-KS(\sigma, t_j) + M_0 + \text{Post} \cdot \delta & \leq 0 & \forall (\sigma, t_j) \in \mathcal{D} \\
\bar{T}S(\sigma, t_j) \leq m - 1 & \forall (\sigma, t_j) \in \mathcal{D} \\
M_0 \in \mathbb{N}^m \\
\text{Pre}, \text{Post} \in \mathbb{N}^{m \times n} \\
S(\sigma, t_j) \in \{0,1\}^m
\end{array}
\right.
\]

where

\[
\mathcal{E} = \{(\sigma, t_j) \mid \sigma \in \mathcal{L}, |\sigma| < k, \sigma t_j \in \mathcal{L}\},
\]

\[
\mathcal{D} = \{(\sigma, t_j) \mid \sigma \in \mathcal{L}, |\sigma| < k, \sigma t_j \notin \mathcal{L}\}
\]

and \(K\) is a very large constant.

Constraints (a) are the enabling constraints, i.e., a transition \(t_j\) is enabled at \(M_0 + (\text{Post} - \text{Pre}) \cdot \delta\) if and only if \(M_0 + (\text{Post} - \text{Pre}) \cdot \delta \geq \text{Pre} \cdot \bar{\varepsilon}\).

Constraints (b) and (c) are the disabling constraints: if a transition \(t_j\) is disabled at \(M_0 + (\text{Post} - \text{Pre}) \cdot \delta\) then there exists at least one place \(p\) such that

\[
M_0(p) + (\text{Post}(p, \cdot) - \text{Pre}(p, \cdot)) \cdot \delta \leq \text{Pre}(p, \cdot) \cdot \bar{\varepsilon} - 1.
\]

Indeed, by constraint (c) at least one entry of \(S(\sigma, t_j)\) is null, thus eq. (2) holds for at least one \(p \in P\). On the contrary, no constraint is given for the other places to which it corresponds a non null entry of \(S(\sigma, t_j)\) because in this case constraint (b) is redundant.

In general the set (1) is not a singleton, thus there exists more than one Petri net system \((N, M_0)\) such that \(L_k(N, M_0) = \mathcal{L}\). To select one among these Petri net systems we choose a given performance index and solving an appropriate IPP we determine a Petri net system that minimizes the considered performance index\(^2\). In particular, if \(f(M_0, \text{Pre}, \text{Post})\) is the considered performance index, an identification problem can be formally stated as follows.

**Problem 3.3** Let us consider the identification problem (3.1) and let \(f(M_0, \text{Pre}, \text{Post})\) be a given performance index. The solution to the identification problem (3.1) that minimizes \(f(M_0, \text{Pre}, \text{Post})\) can be computed by solving the following IPP

\[
\begin{align*}
\min & \quad f(M_0, \text{Pre}, \text{Post}) \\
\text{s.t.} & \quad \mathcal{G}(m, T, \mathcal{L}).
\end{align*}
\]

3.1. Complexity of IPP (3)

Let \(n\) be the cardinality of \(T\), \(k\) the length of the longest string in \(\mathcal{L}\), and \(\nu_r\) (for \(r = 0, \ldots, k\)) the number of strings in \(\mathcal{L}\) of length \(r\).

Then the constraint set (1) contains \(\sum_{r=1}^{k} \nu_r\) constraints of type (a) and \(\sum_{r=0}^{k-1} (\nu_r - \nu_{r+1})\) constraints of type (b) and of type (c). The total number of scalar constraints is thus:

\[
m \left( \sum_{r=1}^{k} \nu_r \right) + (m + 1) \left( \sum_{r=0}^{k-1} (\nu_r - \nu_{r+1}) \right).
\]

The total number of unknowns is

\[
u = m + 2(m \times n) + m \left( \sum_{r=0}^{k-1} (\nu_r - \nu_{r+1}) \right).
\]

Note that given a value of \(k\) and of \(n\), it is possible to find a worst case bound for \(\rho = \sum_{r=0}^{k-1} (\nu_r - \nu_{r+1})\). In fact, it holds:

\[
\rho = \sum_{r=0}^{k-1} (\nu_r - \nu_{r+1}) = \nu_0 + (n - 1) \left( \sum_{r=1}^{k-1} \nu_r \right) - \nu_k = n + (n - 1) \left( \sum_{r=1}^{k-1} \nu_r \right) - \nu_k.
\]

This expression is maximized if we assume \(\nu_k = 0\) while all other \(\nu_r\) must take the largest value, i.e., \(\nu_r = n^r\). Hence we have \(\rho = n + (n - 1) (n + \cdots + n^{k-1}) = n^k\), and the total number of unknowns in the worst case is

\[
u = m + 2(m \times n) + mn^k = m(1 + 2n + n^k) = \mathcal{O}(m n^k),
\]

i.e., it has exponential complexity with respect to \(k\).

\(^1\)A language \(\mathcal{L}\) is said to be prefix-closed if for any string \(\sigma \in \mathcal{L}\), all prefixes of \(\sigma\) are in \(\mathcal{L}\).

\(^2\)Clearly, also in this case the solution may be not unique.
3.2. Optimizing the number of places

If we assume the number \( m \) of places is not given, but it is only known to be less or equal to a given value \( m_0 \), the identification problem 3.1 can be reformulated as follows.

**Problem 3.4** Let us consider an identification problem in the form 3.1 where \( m \) is only known to be less or equal to a given value \( m_0 \), and let \( f(m, M_0, \text{Pre}, \text{Post}) \) be a given performance index. The solution to the identification problem that minimizes \( f(m, M_0, \text{Pre}, \text{Post}) \) with the smallest number of places of \( m \) can be computed solving the following nonlinear IPP

\[
\begin{align*}
\min_{m \leq m_0} & \quad f(m, M_0, \text{Pre}, \text{Post}) \\
\text{s.t.} & \quad G(m, T, \mathcal{L}).
\end{align*}
\]

A trivial solution to the above identification problem 3.4 consists in solving IPP of the form (3) for increasing values of \( m \), until a feasible solution is obtained. Alternatively, the following result can be used.

**Theorem 3.5** [2] Solving the identification problem 3.4 is equivalent to solving the following IPP:

\[
\begin{align*}
\min & \quad J = \bar{K} \cdot \bar{1}_m^T \bar{z} + f(m, M_0, \text{Pre}, \text{Post}) \\
\text{s.t.} & \quad G(m, T, \mathcal{L}) \\
& \quad \bar{K} \cdot \bar{z} = \text{Pre} \cdot \bar{1}_n - \text{Post} \cdot \bar{1}_n \geq \bar{0}_m \\
& \quad z_i \leq z_i, \quad i = 1, \ldots, m - 1 \\
& \quad \bar{z} \in \{0, 1\}^m
\end{align*}
\]

for a sufficiently large constant \( \bar{K} \).

In particular, let us denote as \( \bar{z}^*, M_0^*, \text{Pre}^* \) and \( \text{Post}^* \) the solution of (5), and let \( m^* \) be the number of nonzero components of \( \bar{z}^* \).

Let \( M_0^* \) be the vector obtained from \( \bar{M}_0^* \) by only keeping its first \( m^* \) components. Analogously, let \( \text{Pre}^* \) and \( \text{Post}^* \) be the matrices obtained from \( \text{Pre}^0 \) and \( \text{Post}^0 \), respectively, by only keeping their first \( m^* \) rows.

Then, \( m^*, M_0^*, \text{Pre}^*, \text{Post}^* \) is a solution of the identification problem 3.4.

3.3. Identification from the reachability graph

As discussed in detail in [1] the above procedure can also be used to solve identification problems starting from the reachability graph of the net system, that is represented as a finite state automaton \( G \), provided that the net system is bounded.

To this aim, it is sufficient to define \( \mathcal{L} \) as the set of sequences that are generated by the automaton without passing through a cycle. Then, we need to impose as additional constraints the set of minimal T-invariants \( \Gamma_{\text{min}}(G) \) [1]. More precisely, for any minimal T-invariant \( \bar{y} \in \Gamma_{\text{min}}(G) \) we have to impose an additional constraint of the form

\[
(\text{Post} - \text{Pre}) \bar{y}^T = \bar{0}_m^T.
\]

4. Numerical simulations

Let us consider the Petri net system in Figure 1 consisting of \( 2(q+1) \) places and \( 2q \) transitions. It models a sender-receiver process. In particular, places \( p_i \), with \( i = 1, \ldots, q \), model the sender process; places \( q_i \), with \( i = q+1, \ldots, 2q \), model the receiver process, and places \( p_n \), with \( i = 2q+1, 2q+2 \), correspond to the communication channels between the two processes. When place \( p_1 \) is marked (as in Figure 1) the sender is ready to transmit a message. After the firing of \( t_1 \) the message is in the channel, ready to be received, provided that the receiver is also ready, namely that place \( q_{p+1} \) is marked. Now, transitions \( t_2, \ldots, t_{q-1}, t_{q+1}, \ldots, t_{2q} \) can fire. In particular the firing of \( t_{2q} \) corresponds to the acknowledgment from the receiver. Finally, the receipt from the sender is modeled by transition \( t_q \).

In this section we present the results of various identification problems carried out considering different values of \( q \), namely \( q = 2, \ldots, 6 \). In particular, our goal here is that of synthesizing a net system that generates the same language of the net system in Figure 1.

Note that for sake of brevity we do not report here the language \( \mathcal{L} = L_k(N, M_0) \), but we limit to observe that \( k = 2q \). Moreover, in order to obtain a net system that generates exactly the same language we also need to impose the minimal T-invariant \( \bar{y} = \bar{1}_n \).

Note that we do not need to impose the P-invariants corresponding to the sender and the receiver. In fact, our requirement here is not that of obtaining exactly the net system in Figure 1, but a net system that generates the same language.

For any \( q \) we considered two different cases. In particular, using the notation of IPP (5), we assume:

\[
\begin{align*}
\text{(C1)} & \quad J = \bar{K} \cdot \bar{1}_m^T \bar{z} + \bar{1}_m^T \cdot M_0 + \bar{1}_m^T \cdot (\text{Pre} + \text{Post}) \cdot \bar{1}_n, \\
\text{(C2)} & \quad J = \bar{1}_m^T \cdot M_0.
\end{align*}
\]

Therefore, in case (C1) we assume that the number of places is not known a priori, and our goal is that of determining among all the net systems that satisfy the given language specifications, that one who minimizes
<table>
<thead>
<tr>
<th>$q=2$</th>
<th>First admissible solution</th>
<th>Optimal solution (max 1 hour)</th>
<th>% after 1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>0.03 sec (141)</td>
<td>0.03 sec (141)</td>
<td>0%</td>
</tr>
<tr>
<td>C2</td>
<td>&lt; 0.01 sec (113)</td>
<td>&lt; 0.01 sec (113)</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q=3$</th>
<th>First admissible solution</th>
<th>Optimal solution (max 1 hour)</th>
<th>% after 1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt; 4 sec (598)</td>
<td>16.71%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt; 0.6 sec (607)</td>
<td>0,78 sec (2744)</td>
<td>0%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q=4$</th>
<th>First admissible solution</th>
<th>Optimal solution (max 1 hour)</th>
<th>% after 1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt; 29 sec (1706)</td>
<td>59.91%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt; 8 sec (2479)</td>
<td>50.00%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q=5$</th>
<th>First admissible solution</th>
<th>Optimal solution (max 1 hour)</th>
<th>% after 1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt; 200 sec (3855)</td>
<td>75.05%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt; 20 sec (3625)</td>
<td>75.00%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$q=6$</th>
<th>First admissible solution</th>
<th>Optimal solution (max 1 hour)</th>
<th>% after 1 hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>&lt; 65 sec (6949)</td>
<td>97.61%</td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>&lt; 20 sec (6067)</td>
<td>50.00%</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Numerical results

In the third column we reported a measure of the distance (as a percentage) between the optimal solution and the solution computed within an hour. If such a value is equal to 0% it means that the optimal solution has been obtained in the allowed time. As large is the percentage, as far the solution is from the optimum.

From Table 1 we can be easily observe that the optimal solution can be computed within an hour only for $q = 2$ (both in case C1 and C2), and for $q = 3$ (in case C2). In all the other cases the number of constraints was too high, and regardless of the considered performance index, one hour was not enough to determine the optimal solution. In particular, the distance from the optimal solution in quite all cases examined also depend on the considered performance index (case C1 or C2).

Note however that this is not a serious limitation of our procedure because in general, when computing an identification problem, we are mainly interested in determining an admissible solution, rather than an optimal one. As it can be seen by looking at the first column of Table 1, the computational times are very very short also for large values of $q$ if we consider an arbitrary solution, e.g., the first admissible one computed by CPLEX when solving an optimization problem.

Note that in case C1 we assumed that $\bar{m} = 2q + 2$ but we always found out a net with a minor number of places. This is not surprising because we are not imposing the P-invariants relative to the sender and the receiver, thus the structure of the net is different even if the language generated is the same.

### 5. Conclusions and future work

In this paper we investigated the computational complexity of solving Petri nets identification problems starting from the language generated. Our attention here was limited to bounded Petri net systems. Our future work will be that of extending the proposed procedure to unbounded nets. Moreover, we are working on the design of a MATLAB tool that starting from a given automaton $G$, identifies a net system whose coverability graph is isomorphic to $G$, by solving an appropriate ILL using CPLEX.

### 6. Acknowledgements

We thank Mauro Franceschelli for his help in the development of he tool MATLAB for the generation of the IPP (5) in the syntax of CPLEX.

### References


On Reachability Analysis of Multi Agent Nets
Toshiyuki Miyamoto†, Masaki Sakamoto†, and Sadatoshi Kumagai†
†Graduate School of Engineering, Osaka University
Suita 565-0871, Japan
Email: miyamoto@eei.eng.osaka-u.ac.jp

Abstract—Petri nets are known as a modeling language for concurrent and distributed systems. In recent years, various object-oriented Petri nets were proposed, and we are proposing a kind of object-oriented Petri nets, called multi agent nets (MANs). In this paper, we consider the reachability analysis of MANs. We propose an algorithm for generating an abstract state space of a multi agent net, and report results of computational experiments.

1. Introduction

Petri nets[1] are known as a modeling language for concurrent and distributed systems. So far, Petri nets have been extended in various ways to in order to enhance modeling capability. In recent years, object-oriented Petri nets attracts attentions as a modeling language for multi-agent systems[2]. We are proposing a kind of object oriented Petri nets, called multi agent nets (MANs)[3]. MANs are composed of agent nets, and each agent net is an extended coloured Petri net[4]. An agent net has its own color; therefore an agent net could be located on another agent net as a token. This kind of hierarchy is know as a nets-within-nets structure[5]. Agent nets can communicate each other by passing tokens, and an agent net is possible to decide a communication companion or companions dynamically.

We have been developed a software environment[6] to execute multi agent net models, and developed various distributed multi-agent systems by using the MANs[7, 8, 9]. This paper focuses on the reachability problem of MANs. Nested Petri nets are fundamental class of nets-within-nets style hierarchical Petri nets. Lomazova et al. showed that the reachability problem is undecidable for nested Petri nets[10]. Therefore the reachability problem in MANs is undecidable in general. This paper shows an abstract state space construction algorithm for MANs, and reports results of computational experiments.

2. Multi Agent Nets and Software Environment

2.1. Informal Introduction to Multi Agent Nets

This section informally introduces the MAN. A MAN is a set of agent nets, and each agent net is an extended coloured Petri net.

In coloured Petri nets, a set is associated with each place, and the set is called colour set. A token on a place holds a datum, and it is an element in the associated colour set. We call the data token attributes in this paper. In MANs, the colour set can be also associated with the net itself. Therefore, not only tokens but also nets are able to have some data, called net attributes, and they can be accessed (read and write) from anywhere in the same net. Since each agent net has its own colour, an (lower) agent net can be located on places in other (higher) agent nets as a token. In this case, attributes of the lower agent net can be accessed as attributes of the token.

Agent nets can communicate each other by passing tokens. Internal behaviour of agent nets are encapsulated, and an agent net is able to know a state of another agent net by the following two methods: 1) a higher agent net can know net attributes of lower agent nets, and 2) an agent net can know state of other agent nets by communication. A communication companion is designated in the passing token, and an agent net is possible to decide the communication companion or companions dynamically. In MANs, this token passing mechanism and nets-within-nets style hierarchy bind plural agent nets. In modular Petri nets, there exist two kinds of binding methods: the fusion of transitions, and the fusion of places[11]. Our token passing mechanism corresponds to the fusion of places, however the binding structure has not yet defined explicitly in the net structure. This kind of system is called weakly coupled systems[12].

2.2. Software Environment

Figure 1 shows overview of our software environment for MANs. The part shown by broken lines shows execution modules of MANs, and has been developed out previous works[6]. In the environment, an agent net is written as a class of Java. The class library consists of super class of agent nets and execution module. Our environment also supports distributed execution of MANs.

The part shown by solid lines is developed in this research. AbstractSpaceGenerator generates a reachable space for a given MAN. As the name shows, it generates an abstract state space instead of a complete state space, and the state space is abstracted by an abstraction mapping defined in Abstraction Rules. The abstract state space is expressed by using Binary Decision Diagrams (BDDs). ModelChecker checks the specification written by Computation Tree Logic (CTL). Please refer [13] more detail.
about model checking.

3. State Space Construction

3.1. Markings in MANs

For a given agent net $AN$, let $P$ be the set of places, $\Sigma$ be the set of colour sets, and $C : P \cup \{AN.c\} \rightarrow \Sigma$ be the colour function, where $AN.c$ is the net attribute of agent net $AN$. A token element is a 2-tuple $(p, c)$, and $p \in P \cup \{AN.c\}$ and $c \in C(p)$. Let $TE$ be the set of token elements, then a marking of an agent net is a multi-set over $TE$. A marking $M_{AN}$ of agent net $AN$ is denoted by

$$M_{AN} = 1^i(AN.c, c_\text{an}) + \sum_{i=1}^{m} \sum_{j=1}^{n} N_{ij}(p_i, c_j),$$

where $N_{ij} \in \mathbb{Z}^+$, $N_{ij}(p_i, c_j)$ shows that $N_{ij}$ tokens with attributes $c_j$ are located on place $p_i$, and $c_\text{an}$ shows net attributes.

For given markings $M_{AN}$ and $M'_{AN}$, when $N \leq N'$ for all $(p, c) \in M_{AN}$, marking $M_{AN}$ is said to be covered by $M'_{AN}$, and denoted by $M_{AN} \subseteq M'_{AN}$. $M_{AN}$ and $M'_{AN}$ are called equivalent and denoted by $M_{AN} = M'_{AN}$, when $M_{AN} \subseteq M'_{AN}$ and $M'_{AN} \subseteq M_{AN}$.

A MAN is a set of agent nets, and some of them may be in the same class, namely have the same structure. In order to distinguish, each agent net has identifier. Let $AN_i$ be an identifier of agent net $AN_i$, and $M_i$ be its marking, a marking $M_{MAN}$ of the MAN is expressed by

$$M_{MAN} = \sum_{i=1}^{m} M_{AN_i}(AN_i, M_i),$$

where $I D_M$ is the set of identifiers.

Let us consider equivalence relations among markings of MANs. We introduce two equivalence relations afterward.

One considers the identifier explicitly, and the other one disregards.

**Definition 1 (strongly equivalent)** Two markings

$$M_{MAN} = \sum_{i=1}^{m} |ID_M| (AN_i, M_i)$$

$$M'_{MAN} = \sum_{i=1}^{m} |ID_M| (AN_i', M_i')$$

are said to be strongly equivalent when for all $i$, $AN_i = AN_i'$ and $M_i = M_i'$.

**Definition 2 (weakly equivalent)** Two markings

$$M_{MAN} = \sum_{i=1}^{m} |ID_M| (AN_i, M_i)$$

$$M'_{MAN} = \sum_{i=1}^{m} |ID_M| (AN_i', M_i')$$

are said to be weakly equivalent when there exists a bijection $f$ from $I D_M$ to $I D_M'$ such that for all $i$, $f(AN_i) = AN_i$ and $M_i = M_i'$.

3.2. Abstract State Space

Let $M$ be the set of reachable markings from $M_0$ in a multi agent net $MAN$, $R \subseteq M \times M$ be the set of transition relations among markings, and $I = \{M_0\}$ be the set of initial states, then 3-tuple $(M, R, I)$ is a transition system of $MAN$. For the set $AP$ of atomic propositions, let $\lambda : M \rightarrow 2^{AP}$ be a labeling function, then 4-tuple $C = (M, R, I, \lambda)$ is a Kripke structure.

For this $(M, R, I, \lambda)$, consider a structure $A = (A, R, I, T, X)$. $A$ becomes an abstract structure of $C$ when each term is defined as follows:

Let $\phi_1(M), \ldots, \phi_n(M)$ be predicates and each predicate is a mapping from $M$ to $\mathbb{B}$, and define the set $AP$ of atomic propositions by $AP = \{\phi_1, \ldots, \phi_n\}$. Define the set of states $A$ by

$$A = \{(\phi_1(M), \ldots, \phi_n(M)) | M \in M\},$$

and an abstraction function $\beta : M \rightarrow A$ by $\beta(M) = (\phi_1(M), \ldots, \phi_n(M))$. The set of initial states of $A$ is given by $T = \beta(I)$, and define the transition relation $R$ by

$$R = \{\beta(M), \beta(M')|(M, M') \in R\},$$

and let the labeling function $X$ be an identity mapping.

3.3. State Space Construction Algorithm

We construct an abstract state space by using the depth first search method with a finite depth $D$. In order to calculate the next state, we use the execution module of our software environment. It is shown in Algorithms 1 and 2.

In the algorithm, $M.execute(t)$ returns a set of enabled transitions at a marking $M$, and $M.next(t)$ returns a marking obtained by firing a transition $t$ from a marking $M$.
Algorithm 1: generate state space
1: $\overline{I}, \overline{A} \leftarrow \{\beta(M_0)\}$
2: $\overline{R} \leftarrow \emptyset$
3: $\text{Stack} \leftarrow \emptyset$ // Stack for keeping marking trace
4: $SS \leftarrow \emptyset$ // Set of visited markings
5: $\text{search}(M_0)$

Algorithm 2: search($M$)
1: if $M \in SS$ then
2: return
3: end if
4: $b \leftarrow \beta(M)$
5: $\text{Stack.push}(M)$
6: $SS \leftarrow SS \cup \{M\}$
7: if $\text{Stack.size()} < D$ then
8: $TL \leftarrow M.tlist()$
9: for all $t \in TL$ do
10: $M' \leftarrow M.next(t)$
11: $b' \leftarrow \beta(M')$
12: $A' \leftarrow A \cup \{b\}$
13: $\overline{R} \leftarrow \overline{R} \cup \{(b, b')\}$
14: $\text{search}(M')$
15: end for
16: end if
17: $\text{Stack.pop()}$
18: return

4. Computational Experiments

Figure 2 shows two classes of agent nets. We constructed abstract state spaces for three MANs which are composed by sender nets and receiver nets and the number of sender/receiver nets are changed from one to three. Our program is written in Java, and JavaBDD\(^1\) is used for implementing the transition system. Experiments were done on a PC(CPU: Pen.4 3.2GHz, Mem.: 1G, OS: FC3.0, Java: 1.4.2). We prepared a predicate for each place $p$ in \{\(p_1, p_2, p_3, p_5, p_6, p_8, p_9, p_{10}, p_{11}, p_{12}\}\}. The predicate returns true when there exist an agent net in which at least one token is on the place $p$. Figure 3, 4, 5 shows computation time to construct abstract state spaces and the number of state transitions in the space.

At the first line in Algorithm 2, we check whether the marking $M$ has been visited or not. We prepared three methods for marking equivalence checking. The line titled as "without eq. check" shows the result when we did not checked marking equivalence. The lines titled as “Def. 1” and “Def. 2” show results when the strongly equivalent and the weakly equivalent relations are used, respectively. Results show that state space construction time is enormously reduced by using the marking equivalence relations.

The line title as “number of transitions” is the number of state transitions in the abstract space. Lines for every checking method were the same. This shows that the visited marking check does not bind any necessary search. In Fig. 4, the number of state transitions was 264 in both cases where the search depth was 20 and 25, and the state space construction was saturated.

We are using the abstract state space in order to deal with huge state space of coloured Petri nets, however the algorithm holds the set of visited markings. At this point, our algorithm includes inconsistency. As for the checking of visited markings, we tried to use Stack instead of SS. In this case, we could not see any advantage from “without eq. check” when the search depth was at most 25. The problem is left for our future research.

5. Conclusions

This paper proposed an abstract state space construction algorithm for MANs. When we apply the model checking method to coloured Petri nets such as MANs, state space abstraction is useful.

\(^1\)http://javabdd.sourceforge.net/
The proposed algorithm uses the depth first search with finite search depth, and we proposed two definitions about marking equivalence. We have evaluated the following two points by computational experiments: the visited marking checking contributes enormous time reduction on state space construction, and the visited marking checking does not bound any necessary search.

Acknowledgement

This work was supported in part by Grant-in-Aid for Scientific Research (B) (No. 17360198) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References


How to Obtain Coefficients for a Firing Count Vector Expanded by T-Invariants and Particular Solutions in P/T Petri Nets

Tadashi Matsumoto†, Seiichro Moro‡, and Masahiro Osogami†

†Fukui University of Technology, 3-6-1, Gakuen, Fukui, 910-8505, Japan
‡University of Fukui, 3-9-1, Bunkyo, Fukui, 910-8507, Japan
Email: matsu@fukui-it.ac.jp, moro@ppc8100.fukei-u.ac.jp, osogami@fukui-it.ac.jp

Abstract—Petri nets are basic models for concurrent systems and a firing count vector for transitions is one of key concepts to describe and evaluate algebraically their behavior. In this paper, under the conditions that all minimal T-invariants and all particular solutions for state equation are given, it is shown that the expansion coefficients for any specified firing count vector are systematically obtained by using Modified Fourier-Motzkin method as well as Old Fourier-Motzkin method, both of which have been effectively used also for finding T-invariants and particular solutions of state equation.

1. Introduction

A Petri net is a particular kind of directed graph, together with an initial state called the initial markings, M0. The underlying graph of a Petri net is directed, weighted, bipartite graph consisting of two kinds of nodes, called places and transitions, where arcs are either from a place to a transition or from a transition to a place. A state or marking in a Petri net is changed according to the firing rules. See[1],[2]. Such Petri nets are effectively used for modeling, analyzing, and verifying many discrete event systems[1],[2].

In this paper, we concern structural analysis based on the linear algebra techniques and the net state equation Ax = b := M_d − M_0, where M_0 and M_d are initial and destination marking vectors, respectively. All generators for T-invariants and all minimal inhomogeneous(i.e., particular) solutions are needed for discussing the feasibility of a group of firing count vectors, x, for the fixed b := M_d − M_0[4,6,7], where any firing count vector is expanded by means of T-invariant generators and particular solutions[7]. However, it is difficult, in general, to find the non-negative rational/integer scalar expansion-coefficients. In this paper, we consider how to find systematically those coefficients through the use of well-known Fourier-Motzkin method[3] and the modified one[5].

2. Preliminaries

Let \( Z^{m\times n}(Z^{m\times n}_{++}) \), resp. be the set of \( m \times n \) matrices with integer (nonnegative integer including zero, resp.) elements in this paper. The \( i \)-th element of a vector \( x \) is denoted by \( x(i) \). For two vectors \( x \) and \( y, x > y \) means that \( x(i) > y(i) \) for each \( i \), \( x \geq y \) means that \( x(i) \geq y(i) \) for each, \( i \), and \( x \geq y \) means that \( x(i) \geq y(i) \) for some \( i \). Note that there exists the difference between \( x \geq y \) and \( x \geq y \) in this paper as in Ref.[1], where \( x \geq y \) includes \( x = y \), but \( x \neq y \) does not.

\[ Ax = b \] state equation of a Petri net, where \( A \in Z^{m\times n} \) is an incidence matrix, \( b := M_d − M_0 \in Z^{m\times 1} \) is a difference vector between a destination marking \( M_d \in Z^{m\times 1} \) and an initial marking \( M_0 \in Z^{m\times 1} \) on each place \( p \), and \( x \in Z^{n\times 1} \) is a nonnegative integer firing count vector. Note also that \( m(n, \text{resp.}) \) is the finite number of places (transitions, resp.).

Minimal solution vector for \( Ax = b \): a solution vector \( x \in Z^{m\times 1} \) is said to be minimal if there is no other solution vector \( y \in Z^{m\times 1} \) s.t. \( x(i) \geq y(i) \) for each \( i \) in \( Ax = b[1] \).

Minimal T-invariant; \( U_4 := \{ u^{(5)}_4 \in Z^{m\times 1} \} \) nonnegative integer and minimal homogeneous solutions, \( x = u^{(5)}_4 \) of \( Ax = b := M_d − M_0 \in Z^{m\times 1} \) and \( b = 0^{m\times 1}, i = 1,2,\cdots,l_5 \), where \( u^{(5)}_4 \) is also called a minimal T-invariant.

Minimal particular solution; \( V_5 := \{ v^{(5)}_5 \in Z^{m\times 1} \} \) nonnegative integer and minimal inhomogeneous solutions, \( x = v^{(5)}_5 \) of \( Ax = b := M_d − M_0 \in Z^{m\times 1} \) and \( b \neq 0^{m\times 1}, j = 1,2,\cdots,k_5 \), where \( v^{(5)}_5 \) is also called a minimal (and non-negative integer) particular solution.

Firing count vector; \( X := \{ x \in Z^{m\times 1} \} \) nonnegative integer firing count vectors, \( x \), of \( Ax = b := M_d − M_0 \in Z^{m\times 1} \) and \( b \neq 0^{m\times 1} \) or \( = 0^{m\times 1} \), where \( M_0, M_d \in Z^{m\times 1} \).

Fundamental particular solution; a minimal (and non-negative integer) particular solution \( x \in Z^{m\times 1} \) of \( Ax = b := M_d − M_0 \neq 0^{m\times 1} \) is called a fundamental particular solution if \( x \) is never expressed by convex combination of other minimal particular solutions of \( Ax = b \neq 0^{m\times 1} \). We use the next notation; \( V_4 := \{ v^{(4)}_4 \in Z^{m\times 1} \} \) fundamental particular solutions, \( j = 1,2,\cdots,k_4 \) and then \( V_5 \supseteq V_4(k_5 \geq k_4) \) or \( V_5 = V_4 \setminus V_4 \).

Minimal support T-invariant; the set of elements corresponding to nonzero elements in a T-invariant \( x \geq 0^{m\times 1} \) s.t. \( Ax = 0^{m\times 1} \) is called the support of a T-invariant. A support is said to be minimal if no proper nonempty subset of the support is also a support. We call such a T-Invariant a minimal support, or elementary, T-invariant. We use the next notation for the set of minimal support T-invariants; \( U_4 := \{ u^{(4)}_i \in Z^{m\times 1} \} \) minimal support T-invariants, \( i = 1,2,\cdots,l_4 \) and then \( U_5 \supseteq U_4(l_5 \geq l_4) \) or \( U_5 = (U_4, U_5 \setminus U_4) \).

407
Note that \( U_4 \) is obtained by Old FM method (see Appendix A) and \( U_5 \) by Modified FM method (see Appendix B). Moreover, \( V_5 \) is obtained by Modified FM method for the augmented system of \( Ax = b \) (see Appendix C).

Let \( Q_{\text{mxx}}^+ \) be the set of \( m \times n \) matrices with nonnegative rational elements including zero in this paper.

**Example 1:**

Consider a Petri net shown in Fig.1, where unity weight on each arc is omitted and a black dot (a small white circle, resp.) on place is an initial (a destination, resp.) marking or token. We have \( A \in \mathbb{Z}^{3 \times 5} \) and \( b \in \mathbb{Z}^{3 \times 1} \) for Fig.1 as follows:

\[
A = \begin{bmatrix}
-1 & -1 & 0 & 0 & 1 \\
1 & 1 & -1 & -1 & 0 \\
0 & 0 & 1 & 1 & -1
\end{bmatrix},
\]

\[
b = M_d - M_0 = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 \\
0 & 0 & 1 & 1 & 0
\end{bmatrix}.
\]

For this simple example, we have \( U_5 = \{u_1^{(5)}, u_2^{(5)}, u_3^{(5)}, u_4^{(5)}\} = U_4 \) and \( V_3 = \{v_1^{(5)}, \ldots, v_6^{(5)}\} \), where

\[
u_1^{(5)} = u_1^{(4)} = (10101)^T, \quad v_1^{(5)} = v_1^{(4)} = (10200)^T
\]

\[
u_2^{(5)} = u_2^{(4)} = (01011)^T, \quad v_2^{(5)} = v_2^{(4)} = (10020)^T
\]

\[
u_3^{(5)} = u_3^{(4)} = (01101)^T, \quad v_3^{(5)} = v_3^{(4)} = (01200)^T
\]

\[
u_4^{(5)} = u_4^{(4)} = (01011)^T, \quad v_4^{(5)} = v_4^{(4)} = (01020)^T
\]

\[
u_5^{(5)} = (10110)^T = \frac{1}{2} v_1^{(5)} + \frac{1}{2} v_2^{(5)}
\]

\[
u_6^{(5)} = (01110)^T = \frac{1}{2} v_4^{(5)} + \frac{1}{2} v_6^{(5)}
\]

Note that \( u_1^{(5)}, u_2^{(5)}, u_3^{(5)}, u_4^{(5)} \) and \( u_5^{(5)} \) are minimal (and minimal support) T-invariants, and \( v_1^{(5)}, v_2^{(5)}, v_3^{(5)}, v_4^{(5)} \) and \( v_5^{(5)} \) are fundamental (and minimal) particular solutions, and \( v_5^{(5)} \) and \( v_6^{(5)} \) are minimal, but not fundamental, particular solutions.

**3. An Arbitrary Firing Count Vector by Means of T-Invariants and Particular Solutions**

A firing count vector \( x \in \mathbb{Z}_{\text{mxx}}^+ \) is expressed by using \( u_i^{(4)} \in U_4 = \{\text{the set of fundamental particular solutions}\} \) and \( v_j^{(5)} \in V_5 \) (the set of fundamental particular solutions) as follows[7]:

\[
x = \sum_{i=1}^{k_4} a_i^{(4)} u_i^{(4)} + \sum_{j=1}^{k_5} \beta_j^{(5)} v_j^{(5)}, \quad \sum_{j=1}^{k_5} \beta_j^{(5)} = 1,
\]

where \( a_i^{(4)} \in \mathbb{Q} \) and \( \beta_j^{(5)} \in \mathbb{Q}^1 \). We call eq.(1) the level 4 expression in this paper. Moreover we have another expression for \( x \in \mathbb{Z}_{\text{mxx}}^+ \) if we use \( U_5 = \{U_4, U_5\} \) where \( U_5 \) is the set of minimal T-invariants and \( V_5 = \{V_4, V_5\} \) is the set of minimal particular solutions as follows[7]:

\[
x = \sum_{i=1}^{k_4} a_i^{(5)} u_i^{(5)} + \sum_{j=1}^{k_5} \beta_j^{(5)} v_j^{(5)},
\]

where \( \sum_{j=1}^{k_5} \beta_j^{(5)} = 1, I_5 = |U_5|, k_5 = |V_5|, \) and \( a_i^{(5)}, \beta_j^{(5)} \in \mathbb{Q}^1 \). Then eq.(2) is rewritten as

\[
x = \sum_{i=1}^{k_4} a_i^{(5)} u_i^{(5)} + v_j^{(5)},
\]

where \( v_j^{(5)} \in V_5 \) and \( a_i^{(5)} \in \mathbb{Q}^1 \). We call eq.(2) or (3) as the level 5 expression in this paper.

**4. How to Find Coefficients of Eqs.(1) to (3)**

**4.1. Problems to Be Considered in This Paper**

Problems considered in this paper are as follows[6].

(1) Find coefficients \( a_i^{(4)}, \beta_j^{(5)} \in \mathbb{Q}^1 \) of eq.(1) when a firing count vector \( x \in \mathbb{Z}_{\text{mxx}}^+ \) s.t. \( Ax = b \) is specified under the given \( U_4 \) and \( V_4 \), where \( \sum_{j=1}^{k_5} \beta_j^{(4)} = 1 \) is satisfied.

(2) Find coefficients \( a_i^{(5)}, \beta_j^{(5)} \in \mathbb{Z}_{\text{mxx}}^+ \) of eq.(2) when a firing count vector \( x \in \mathbb{Z}_{\text{mxx}}^+ \) s.t. \( Ax = b \) is specified under the given \( U_5 \) and \( V_5 \), where the only one coefficient \( \beta_j^{(5)} = 1 \) is nonzero and the others are zero.

(3) Find coefficients \( a_i^{(5)} \in \mathbb{Z}_{\text{mxx}}^+ \) of eq.(3) when \( x \in \mathbb{Z}_{\text{mxx}}^+ \) and \( a_j^{(5)} \in V_5 \) are specified under the given \( U_5 \) and \( V_5 \).

**4.2. Formulations for Three Problems in §4.1**

(1) Problem (1) in §4.1 is expressed through eq.(1) with respect to the unknown coefficients as

\[
[\begin{array}{cccccccc}
u_1^{(5)} & \ldots & \ldots & \ldots & \nu_4^{(5)} & \ldots & \ldots & \nu_6^{(5)} \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1
\end{array}] = x \in \mathbb{Z}^{(n+1) \times 1}
\]

where \( a_i^{(4)} \in \mathbb{Q}^1 \) and \( \beta_j^{(5)} \in \mathbb{Q}^1 \).

(2) Problem (2) in §4.1 is formulated through eq.(2) as
(5)

\[
\begin{bmatrix}
\alpha^{(5)}_1, \alpha^{(5)}_2, \ldots, \alpha^{(5)}_n
\end{bmatrix}
= \begin{bmatrix}
x
1
\end{bmatrix}
\in \mathbb{Z}^{(n+1)\times 1},
\]

where \(\alpha^{(5)} = (\alpha^{(5)}_1, \alpha^{(5)}_2, \ldots, \alpha^{(5)}_n)^T \in \mathbb{Z}^{n\times 1}\) and \(\beta^{(5)} = (\beta^{(5)}_1, \ldots, \beta^{(5)}_n)^T \in \mathbb{Z}^{n\times 1}\). Note that \(\beta^{(5)}\) is the only one nonzero and unity element.

(3) Problem (3) in §4.1 is rewritten follows:

\[
[u^{(5)}_1, \ldots, u^{(5)}_n] \alpha^{(5)} = [x - v^{(5)}_j] \in \mathbb{Z}^{n\times 1},
\]

where \(\alpha^{(5)} \in \mathbb{Z}^{n\times 1}\). Note that each equation of eqs.(4) to (6) is the type \(Ax = b\). Then we can apply Appendix C(i.e., Appendices A and B) to solving of eqs.(4) to (6).

### 4.3. Main Properties for Three Problems

(1) Finding the unknown coefficients of eqs.(4) to (6) is to search nonnegative rational or integer particular solutions.

(2) In “the type \(Ax = b\)” of eqs.(4) to (6), each elements of the matrix \(A\) and the matrix \(b\) is nonnegative integer. Then, the augmented matrix \(\tilde{A} = [A, -b]\) has negative integer elements at only the last column. See Appendix C. From this fact and Old FM Method or Modified FM method for solving \(\tilde{A}x = 0\) in Appendices, the above property (1) is proved.

(3) Modified FM method and Old FM method are useful for the above property (1). The former is perfect, but the latter is not perfect because Old FM method gives only a part of fundamental particular solutions at the most.

(4) In Problem 3 in §4.1, i.e., eq.(6), it happens for eq.(6) to have no solution because it happens that the selected \(v^{(5)}_j \in V_5\) cannot cover the specified \(x \in \mathbb{Z}^{n+1}\). However, note that eq.(6) can have the solution for the same specified \(x \in \mathbb{Z}^{n+1}\) by selecting another \(v^{(5)}_j \in V_5\) one by one.

### 5. Example

Example 2(example the example in §4.1) From the results of Example 1, we have \(U_4 = U_3 = [u^{(5)}_1, u^{(5)}_2, u^{(5)}_3, u^{(5)}_4], V_4 = [v^{(5)}_1, v^{(5)}_2, v^{(5)}_3, v^{(5)}_4]\), and \(V_5 = [v^{(5)}_1, v^{(5)}_2, v^{(5)}_3, v^{(5)}_4]\). Now, for example, specify the firing count vector \(x = (4, 2, 3, 4, 5)^T\) and \(v^{(5)}_5 = (1, 0, 1, 1, 1)^T \in V_5 \setminus V_4\). Then find all possible \((\alpha^{(5)}_1, \ldots, \alpha^{(5)}_4)\). We have \(u^{(5)}_1, u^{(5)}_2, u^{(5)}_3, u^{(5)}_4 \in [a^{(5)}_1, \ldots, a^{(5)}_4]^T = x - v^{(5)}_j\), that is,

\[
\begin{bmatrix}
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
1 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 \\
1 & 1 & 1 & 1
\end{bmatrix}
\begin{bmatrix}
a^{(5)}_1 \\
a^{(5)}_2 \\
a^{(5)}_3 \\
a^{(5)}_4
\end{bmatrix}
= \begin{bmatrix}
3 \\
2 \\
2 \\
3 \\
5
\end{bmatrix}
\]

The augmented matrix \(\tilde{A}\) is

\[
\tilde{A} = \begin{bmatrix}
1 & 1 & 0 & 0 & -3 \\
0 & 0 & 1 & 1 & -2 \\
1 & 0 & 1 & 0 & -2 \\
0 & 1 & 0 & 1 & -3 \\
1 & 1 & 1 & 1 & -5
\end{bmatrix}
\in \mathbb{Z}^{5\times 5}.
\]

From Old FM method, \((\alpha^{(5)}_1, \ldots, \alpha^{(5)}_4)^T = (0320)^T\) and \((2102)^T\) are obtained. From Modified FM method, \((1211)^T\) as well as the above two are obtained, where note that \((1211)^T = \frac{1}{2}(0320)^T + \frac{1}{2}(2102)^T\).

### 6. Conclusions

For any firing count vector \(x \in \mathbb{Z}^{n+1}\) expressed by means of minimal support or minimal \(T\)-invariants and fundamental or minimal particular solutions for state equation in P/T Petri nets, how to find systematically nonnegative rational or integer coefficients has been shown through Modified Fourier-Motzkin method. Although Modified Fourier-Motzkin method as well as the old one is effectively used to have \(T\)-invariants and particular solutions, it is noted that the same algorithms can be applied to have the expansion coefficients.

These results are useful for the reachability analysis and the scheduling for the fixed initial and destination markings.

### Acknowledgments

This research is partly supported by FUT Research Promotion Fund. Many thanks also for the late Miss Takata.

### References


Appendix A: Old Fourier-Motzkin Method

The method for obtaining the set of T-invariants is called Old Fourier-Motzkin method[3] in this paper. However, it is noted that this method cannot always obtain minimal, but not minimal support, T-invariants \( u_i^{(5)} \in U_3 \setminus U_4 \). Old Fourier-Motzkin method is as follows.[3],[5]

Old FM Method

Input: Incidence matrix \( A \in \mathbb{Z}^{m \times n} \), \( m \), and \( n \).
Output: The set of T-invariants including all minimal support T-invariants.

Initialization: The matrix \( B \) is constructed by adjoining the identity matrix \( E^{mn} \) to the bottom of the incidence matrix \( A \in \mathbb{Z}^{m \times n} \), with \( B = [A^T, E]^T \in \mathbb{Z}^{(m+n) \times n} \).

The following operations a), b) are repeated from \( i = 1 \) to \( m = |P| \), where \( |P| \) means the cardinality of the place set \( P \).

a) Add to the matrix \( B \) all the columns which are linear combinations of pairs of columns of \( B \) and which annul the \( i \)-th row of \( B \).

b) Eliminate from \( B \) the columns in which the \( i \)-th element is nonzero.

When this algorithm finished, each column of the submatrix \( C \in \mathbb{Z}^{n \times n} \) which is obtained by deleting the rows from the first to the \( m \)-th from the final outputted matrix \( B \in \mathbb{Z}^{m \times n} \) is a T-invariant. However, in general, this submatrix \( C \) includes also non-minimal-support T-invariants. Therefore if the following operation c) is added and applied to \( C \), minimal support T-invariants are only obtained.

c) Each column vector \( u_i \in \mathbb{Z}^{n \times 1} \) which satisfies the rank condition \( q(u_i) = \text{rank} A'(u_i) + 2 \) is removed from the submatrix \( C = [u_i] \in \mathbb{Z}^{n \times n} \). Here, \( q(u_i) \) is the number of nonzero elements of \( u_i \in \mathbb{Z}_+^{n \times 1} \) for \( A u_i = 0^{mn \times 1} \) and \( A'(u_i) \) is composed of the columns of \( A \), of which columns are corresponding to nonzero elements of \( u_i \in \mathbb{Z}_+^{n \times 1} \).

Appendix B: Modified Fourier-Motzkin Method

Note that \( u_i^{(4)} \in U_4 \) is always obtained, but \( u_i^{(5)} \in U_3 \setminus U_4 \) cannot be always obtained by applying Old Fourier-Motzkin method to the incidence matrix \( A \in \mathbb{Z}^{m \times n} \). Therefore, we must modify Old Fourier-Motzkin method as follows to obtain always all of minimal T-invariants for \( Ax = b \).

<Modified FM Method>

Input: Incidence matrix \( A \in \mathbb{Z}^{m \times n} \), \( m \), and \( n \).
Output: All of minimal T-invariants.

Initialization: The matrix \( B \) is constructed by adjoining the identity matrix \( E^{mn} \) to the bottom of the incidence matrix \( A \in \mathbb{Z}^{m \times n} \), with \( B = [A^T, E]^T \in \mathbb{Z}^{(m+n) \times n} \).
Step0: \( i = 1 \).
Step1: Select the \( i \)-th row of \( B \). If the \( i \)-th row has no nonzero element, then \( i = i + 1 \) and go to Step2. If the \( i \)-th row has at least one nonzero element, then go to Step3.
Step2: If \( i \leq m \) is satisfied, go to Step1, otherwise go to Step8.
Step3: If the \( i \)-th row of \( B \) has at least one pair of positive and negative elements, go to Step4, otherwise go to Step7.
Step4: Aiming the \( i \)-th row of \( B \) (i.e., the old matrix), add directly the \( j \)-th column to the \( k \)-th column, where the \( (i, j) \) element is positive and the \( (i, k) \) element is negative. Apply the minimal vector criterion to the above new column vector and the column vectors each of which has the zero \( i \)-th element on the old matrix \( B \). Adjoin all the remained columns after this criterion to the old matrix \( B \). Then call this new matrix as \( B \) again. Then go to Step5.
Step5: If the \( i \)-th element of all the adjoined column vectors of the new matrix \( B \) is zero, go to Step7, otherwise go to Step6.
Step6: Repeat Step4 to the matrix \( B \). However, the \( (j, k) \) pair should be always new. Then, go to Step5.
Step7: Delete, from \( B \), all the columns each of which has nonzero element on the \( i \)-th row of \( B \). Now, let us call the new matrix as \( B \) again. Then set \( i = i + 1 \) and go to Step2.
Step8: Each column of the submatrix which is obtained by deleting the rows of the first to the \( m \)-th from \( B \) is a minimal nonnegative integer solution \( u_i^{(5)} \) for \( Ax = 0^{mn \times 1} \).

<The Minimal Vector Criterion to \( B \in \mathbb{Z}^{m \times (n+q)} \)>

Aiming the \( i \)-th row of \( B \), let \( b^i \in \mathbb{Z}^{m \times (n+1)} \) and \( b^j \in \mathbb{Z}^{m \times (n+1)} \) be two column vectors of \( B \), where \( k \neq l \) and the \( i \)-th element of \( b^k \) is zero. Let \( \tilde{b}^k \in \mathbb{Z}^{m \times 1} \) and \( \tilde{b}^j \in \mathbb{Z}^{m \times 1} \) be two column vectors which are obtained by deleting the upper \( m \) elements from \( b^k \) and \( b^j \), respectively. If \( \tilde{b}^j \leq \tilde{b}^i \), then delete \( b^l \) from \( B \). Repeat the above for all possible \( b^k \). If \( b^k = b^j (\alpha 
eq \beta) \), then delete \( b^j \) from \( B \).

Appendix C: Augmented Incidence Matrix

1) If we apply Modified FM method to the augmented incidence matrix \( \tilde{A} = [A, -b] \in \mathbb{Z}^{(m+n+1) \times 1} \) for state equation \( Ax + b = 0^{mn \times 1} \), we obtain all minimal T-invariants \( \tilde{u}_i \in \mathbb{Z}_+^{n+1 \times 1} \) of \( \tilde{A} \) as follows[5].
2) Now, we can find the set \( \{U_5, V_5\} \) at level 5 for \( x \in \mathbb{Z}_+^{n \times 1} \) of \( Ax = b \) from \( \tilde{u}_i \in \tilde{U}_5 := \{\tilde{u}_i \in \mathbb{Z}_+^{n+1 \times 1} \} \) nonnegative integer minimal T-invariants of \( \tilde{A} = 0^{mn \times 1} \) as follows[5].

(1) If \( \tilde{u}_i(n+1) = 0^1 \) on \( \tilde{u}_i \in \tilde{U}_5 \), then \( u_i^{(5)} = (\tilde{u}_i(1), \ldots, \tilde{u}_i(n))^T \in \tilde{U}_5 \).
(2) If \( \tilde{u}_j(n+1) = 1^1 \) on \( \tilde{u}_j \in \tilde{U}_5 \), then \( v_j^{(5)} = (\tilde{u}_j(1), \ldots, \tilde{u}_j(n))^T \in V_5 \).
(3) If \( \tilde{u}_k(n+1) > 1^1 \) on \( \tilde{u}_k \in \tilde{U}_5 \), then \( x_k = (\tilde{u}_k(1), \ldots, \tilde{u}_k(n))^T \in \mathbb{Z}_+^{n \times 1} \) is not a solution of \( Ax = b \).
Computation of Controllable Sublanguages for Unbounded Petri Nets Using Their Approximation Models

Shigemasa Takai† and Yongming Bai‡

†Department of Information Science, Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto, 606-8585 JAPAN
‡Graduate School of Systems Engineering, Wakayama University
Sakaedani, Wakayama, 640-8510 JAPAN
Email: takai@kit.ac.jp

Abstract—We study computation of a controllable sublanguage of a given non-prefix-closed regular specification language for an unbounded Petri net. We approximate the generated language of the unbounded Petri net by a regular language, and compute the supremal controllable sublanguage of the specification language with respect to the regular language approximation. This computed language is a controllable sublanguage with respect to the original generated language of the unbounded Petri net, but is not necessarily the supremal one. We then present a sufficient condition under which the computed controllable sublanguage is the supremal controllable sublanguage with respect to the original generated language of the unbounded Petri net.

1. Introduction

Computing the supremal controllable sublanguage of a given specification language is a fundamental problem in supervisory control of discrete event systems (DESs) [8], [10], [1]. The supremal controllable sublanguage is effectively computable when both the generated language of the system and the specification language are regular [10]. A maximally permissive supervisor is implemented by a finite automaton that marks the supremal controllable sublanguage.

In this paper, we assume that a DES to be controlled is modeled by an unbounded Petri net [7] and a given specification language is regular. Since the generated language of an unbounded Petri net is non-regular in general, the results for regular languages are not applicable. A maximally permissive supervisor for a Petri net is synthesized in [6] and [5] under the assumption that a specification language is prefix-closed. Due to this assumption, the results of [6] and [5] can be used only when blocking is not an issue. To our knowledge, computation of the supremal controllable sublanguage of a non-prefix-closed regular specification language for Petri nets remains open. Some difficulties in computing the supremal controllable sublanguage for Petri nets are discussed in [3] and [4]. A main purpose of this paper is to compute a controllable sublanguage of a non-prefix-closed regular specification language for an unbounded Petri net. We first construct a finite automaton whose generated language approximates the generated language of the unbounded Petri net. The constructed finite automaton is a modified version of the coverability graph [9], and is parameterized by a nonnegative integer. We then compute the supremal controllable sublanguage of the regular specification language with respect to the regular language approximation. This computed language is a controllable sublanguage with respect to the original generated language of the unbounded Petri net. As the parameter of the constructed finite automaton is increased, the computed controllable sublanguage becomes larger. However, the computed controllable sublanguage is not necessarily the supremal one. We present a sufficient condition under which the computed sublanguage is the supremal controllable sublanguage with respect to the original generated language of the unbounded Petri net.

2. Preliminaries

Let \( N \) be the set of all nonnegative integers. We consider an unbounded Petri net \( PN = (P, T, A, M_0) \), where \( P \) is the finite set of places, \( T \) is the finite set of transitions, \( A : (P \times T) \cup (T \times P) \rightarrow N \) is a function that specifies the number of arcs between a place and a transition, and \( M_0 \in N^P \) is the initial marking. A transition \( t \in T \) is said to be enabled at a marking \( M \in N^P \) if

\[
\forall p \in P; \ M(p) \geq A(p, t). \tag{1}
\]

The firing of an enabled transition \( t \in T \) at \( M \) yields a new marking \( M' \), which is defined as

\[
\forall p \in P; \ M'(p) = M(p) + A(t, p) - A(p, t).
\]

The notation \( M[t > M'] \) denotes that the firing of \( t \) at \( M \) yields \( M' \). For any markings \( M \) and \( M' \), the notation \( M \preceq M' \) denotes that \( M(p) \leq M'(p) \) for all \( p \in P \).
Let $T^*$ be the set of all finite firing sequences of elements of $T$, including the empty sequence $\varepsilon$. A subset of $T^*$ is called a language (over $T$). A firing sequence $u = t_1t_2t_3 \cdots t_n \in T^*$ is said to be enabled at a marking $M$ if there exist markings $M_i$ ($i = 1, 2, \ldots, n$) such that $M[t_1] > M_1[t_2] > M_2 \cdots M_{n-1}[t_n] > M_n$. The notation $M[u] >$ denotes that $u$ is enabled at $M$, and $M[u] > M'$ denotes that the firing of $u$ at $M$ yields $M'$. Let $M[e] >$ and $M[e] > M$ for any marking $M$. The generated language of $PN$, denoted by $L(PN)$, is defined as

$$L(PN) = \{ u \in T^* \mid M_0[u] > \}.$$  

Let $L \subseteq T^*$ be a language. We denote the set of all prefixes of sequences in $L$ by $\mathcal{L}$, i.e., $\mathcal{L} = \{ u \in T^* \mid \exists v \in T^* \colon uv \in L \}$. $L$ is said to be prefix-closed if $L = \mathcal{L}$.

For the supervisory control purpose [8], the set $T$ is partitioned into two disjoint subsets $T_c$ and $T_u$ of controllable and uncontrollable transitions, respectively. We assume that a control specification for $PN$ is given by a language $K \subseteq L(PN)$. Note that in this paper, we do not assume that $K$ is prefix-closed. A language $L \subseteq L(PN)$ is said to be controllable [8] with respect to $L(PN)$ if

$$TT_u \cap L(PN) \subseteq \mathcal{L}.$$  

We need to compute a controllable sublanguage of $K \subseteq L(PN)$ with respect to $L(PN)$ to synthesize a supervisor that enforces the specification. Let $C(K, L(PN))$ be the set of all controllable sublanguages of $K$ with respect to $L(PN)$. The supremal element of $C(K, L(PN))$ always exists, and is called the supremal controllable sublanguage with respect to $L(PN)$ [10]. We denote this supremal sublanguage by $K^{*}(L(PN))$. The supremal sublanguage $K^{*}(L(PN))$ is effectively computable when both $K$ and $L(PN)$ are regular [10]. However, $L(PN)$ is not necessarily regular for an unbounded Petri net $PN$.

In this paper, we approximate $L(PN)$ by a regular language $L$ such that $L(PN) \subseteq L$, and compute a controllable sublanguage in $C(K, L(PN))$ using the approximation $L$ under the assumption that $K$ is regular.

3. Computation of Regular Language Approximations

In this section, we present an algorithm to construct a finite automaton that approximates the behavior of an unbounded Petri net $PN$.

A finite automaton is a five-tuple $G = (Q, \Sigma, \delta, q_0, Q_m)$, where $Q$ is the finite set of states, $\Sigma$ is the finite set of events, a partial function $\delta : Q \times \Sigma \to Q$ is the transition function, $q_0 \in Q$ is the initial state, and $Q_m \subseteq Q$ is the set of marked states. The function $\delta$ can be generalized to $\delta : Q \times \Sigma^* \to Q$ in the natural way. For any $s \in \Sigma^*$ and $q \in Q$, the notation $\delta(q, s)$ denotes that $\delta(q, s)$ is defined. The generated and marked languages of $G$, denoted by $L(G)$ and $L_m(G)$, respectively, are defined as $L(G) = \{ s \in \Sigma^* \mid \delta(q_0, s) \}$ and $L_m(G) = \{ s \in \Sigma^* \mid \delta(q_0, s) \in Q_m \}$.

Given an unbounded Petri net $PN$, we construct a finite automaton $G^k_{PN}$ as follows:

1. Choose a nonnegative integer $k \in N$, and let

$$N_k = \{ 0, 1, 2, \cdots, k \} \cup \{ \omega \}.$$  

The symbol $\omega$ is introduced to represent “infinity” in a marking, and satisfies $n < \omega$ and $n + \omega = \omega$ for any integer $n$ [7].

2. The transition function $\delta_k : N_k^P \times T \to N_k^P$ is defined as follows. For any $M \in N_k^P$ and $t \in T$, $\delta_k(M, t)$ if and only if (1) holds. If $\delta_k(M, t)$, then the new marking $M' := \delta_k(M, t) \in N_k^P$ is defined as

$$\forall p \in P; \ M'(p) = \begin{cases} N(p), & \text{if } N(p) \leq k \\ \omega, & \text{otherwise}. \end{cases}$$  

where $N(p) = M(p) + A(t, p) - A(t, p)$.

3. A finite automaton $G^k_{PN}$ is defined as

$$G^k_{PN} = (N_k^P, T, \delta_k, M_{k,0}, N_k^P),$$  

where the initial marking $M_{k,0} \in N_k^P$ is given by

$$\forall p \in P; \ M_{k,0}(p) = \begin{cases} M_0(p), & \text{if } M_0(p) \leq k \\ \omega, & \text{otherwise}. \end{cases}$$

We identify $G^k_{PN}$ with its accessible part [2].

Remark 1 In the construction of $G^k_{PN}$, the symbol $\omega$ appears when the number of tokens in a place exceeds $k$. On the other hand, in the coverability graph, $\omega$ appears when a marking strictly covers another marking.

Example 1 We consider an unbounded Petri net $PN$ shown in Fig. 1. The reachability graph of $PN$ is shown in Fig. 2. The finite automaton $G^2_{PN}$ constructed by the above procedure for $k = 2$ is shown in Fig. 3.

The following theorem shows that the behavior of the finite automaton $G^k_{PN}$ includes that of the original unbounded Petri net $PN$.

Theorem 1 For any $k \in N$,

$$L(PN) \subseteq L(G^k_{PN}).$$  

The following theorem shows that as $k$ is increased, the regular language approximation $L(G^k_{PN})$ becomes more accurate.
Figure 1: An unbounded Petri net $PN$. 

$$(1,0,0) \quad (1,0,1) \quad (1,0,2) \quad (1,0,3)$$

$$t_3 \quad t_4 \quad t_3 \quad t_4 \quad \ldots$$

$$(0,1,0) \quad (0,1,1) \quad (0,1,2) \quad (0,1,3)$$

Figure 2: The reachability graph of $PN$. 

Example 2 We consider the unbounded Petri net $PN$ shown in Fig. 1. The finite automaton $G^1_{PN}$ constructed for $k = 1$ is shown in Fig. 4. In this example, $L(G^k_{PN})$ is strictly smaller than $L(G^1_{PN})$.

4. Computation of Controllable Sublanguages Using Regular Language Approximations

Throughout this section, we assume that a specification language $K \subseteq L(PN)$ is regular. Let $G^k_{PN}$ be the finite automaton constructed in the last section. Then, the supremal controllable sublanguage $K^{↑L(G^k_{PN})}$ of $K$ with respect to $L(G^k_{PN})$ is effectively computable. The following proposition presents useful properties of $K^{↑L(G^k_{PN})}$.

**Proposition 1** For any $k \in \mathcal{N}$, 
1. $K^{↑L(G^k_{PN})} \subseteq \mathcal{C}(K, L(PN))$,
2. $K^{↑L(G^k_{PN})} \subseteq K^{↑L(G^{k+1}_{PN})}$.

The first part of Proposition 1 shows that $K^{↑L(G^k_{PN})}$ is a controllable sublanguage of $K$ with respect to $L(PN)$. The second part shows that as $k$ is increased, the controllable sublanguage $K^{↑L(G^k_{PN})}$ becomes larger.

**Remark 2** In the worst case, the number of states of $G^k_{PN}$ is $(k + 2)^{|P|}$. As $k$ is increased, the number of states of $G^k_{PN}$ is also increased. So there is a tradeoff between the permissiveness of the controlled behavior and the computational cost.

The controllable sublanguage $K^{↑L(G^k_{PN})}$ is not necessarily the supremal controllable sublanguage $K^{↑L(PN)}$ with respect to $L(PN)$. The following theorem presents a sufficient condition under which $K^{↑L(G^k_{PN})}$ is equal to $K^{↑L(PN)}$.

**Theorem 3** For any $k \in \mathcal{N}$, if
$$\overline{\mathcal{T}}_u \cap L(G^k_{PN}) \subseteq L(PN),$$
then $K^{↑L(G^k_{PN})} = K^{↑L(PN)}$.

**Remark 3** Since $K$ and $L(G^k_{PN})$ are regular, $\overline{\mathcal{T}}_u \cap L(G^k_{PN})$ is regular. Also, since we consider a free-labeled Petri net $PN$ [7], $L(PN)$ is a P-type Petri net language generated by a deterministic Petri net. By Proposition 3.1 and Lemma 5.1 of [4], we can conclude that it is decidable whether the inclusion relation of (2) holds.
Example 3 We consider the unbounded Petri net $PN$ shown in Fig. 1, and assume that $T_c = \{t_1, t_2\}$ and $T_u = \{t_3, t_4\}$. Let $K \subseteq L(PN)$ be a regular language marked by a finite automaton $G_K$ shown in Fig. 5, where a double circle represents a marked state.

![Figure 5: A finite automaton $G_K$.](image)

We first compute a controllable sublanguage $K^{\dagger, L(G_{PN})}$ by using the finite automaton $G_{PN}$. A finite automaton $G_{K^{\dagger, L(G_{PN})}}$ which marks $K^{\dagger, L(G_{PN})}$ is shown in Fig. 6. We can verify that $K^{\dagger, L(G_{PN})}$ is not the supremal controllable sublanguage with respect to $L(PN)$.

We next compute a controllable sublanguage $K^{\dagger, L(G_{PN})}$ by using the finite automaton $G_{PN}$. A finite automaton $G_{K^{\dagger, L(G_{PN})}}$ which marks $K^{\dagger, L(G_{PN})}$ is shown in Fig. 7. Since $L(G_{PN})$ satisfies (2), $K^{\dagger, L(G_{PN})}$ is equal to the supremal controllable sublanguage $K^{\dagger, L(PN)}$ with respect to $L(PN)$.

![Figure 6: A finite automaton $G_{K^{\dagger, L(G_{PN})}}$.](image)

![Figure 7: A finite automaton $G_{K^{\dagger, L(G_{PN})}}$.](image)

5. Conclusion

In this paper, we have studied computation of a controllable sublanguage of a given non-prefix-closed regular specification language for an unbounded Petri net. We have constructed a finite automaton whose generated language approximates the generated language of the unbounded Petri net, and computed a controllable sublanguage using this finite approximation model. Further, we have presented a sufficient condition under which the computed sublanguage is the supremal controllable sublanguage.

Acknowledgment

This work was supported in part by MEXT under Grant-in-Aid for Scientific Research (No. 18560433).

References


Compositional Controllability Analysis for Partially-Strict Composed DES

Florian Wenck† and Francisco A. C. Silva Neto‡

† Institute of Automation and Computer Control, Ruhr-Universitaet Bochum
Universitaetsstrasse 150, 44780 Bochum, GERMANY
‡ Systems and Automation Department, Federal University of Santa Catarina
Campus Universitário, 88040-900, Florianópolis, BRAZIL.
Email: wenck@atp.rub.de, assisneto@yahoo.com

Abstract—This work addresses large-scale discrete-event systems (DES) whose components run in terms of a partial lockstep. Compositional modelling is required and results in a structured system model. Existing monolithic methods to analyze controllability are based on global unstructured models and thus fail, due to their computational complexity. Compositional analysis remedies this issue, as presented in previous work for master-slave and strict discrete-event system composition. The approach operates without explicitly composing the global model. Rather, compositional analysis retains the given structured model, leading to complexity reduction and computable controllability analysis for systems of industrial size. In this paper the approach is extended to partially-strict composed DES. Necessary and sufficient conditions to verify controllability of such systems in a compositional manner are presented.

1. Introduction

Discrete-event systems (DES) are characterized by a discrete state space and state transitions induced by the asynchronous occurrence of events. Although formal methods do exist to analyze specific properties of DES, such as controllability, these methods require a formal monolithic model of both the system and its specifications. Due to their underlying physical topology, DES of industrial size consist of distributed and coupled components. Specifications are usually defined locally, thus restricting solely a limited amount of components and not the entire system. Analyzing such DES is only feasible using compositional models. Given a compositional model, existing analysis methods require the composition to derive a global model of the system and the specifications. This leads to the well known issue of an exponentially increasing state space. As a direct consequence, such monolithic analysis methods fail, due to their computational complexity.

In [1] and [2] the framework of compositional analysis is introduced for systems whose components are coupled in the sense of a master-slave or a complete lockstep. The approach analyzes significant properties of large-scale DES avoiding the composition, thus leading to complexity reduction. This paper extends the framework using preliminary results from [3] to partially-strict composed DES. Therefore necessary and sufficient conditions are determined to test controllability directly on the structured model of a partially-strict composed DES.

The remainder of the paper is organized as follows. Partially-strict composed DES are discussed in Sec. 2. The problem considered here is stated in Sec. 3. Subsequently, the main results are presented in Sec. 4 and illustrated by an example in Sec. 5. The paper concludes with Sec. 6.

2. Preliminaries

2.1. Notation

The formal Supervisory Control Theory (SCT) presented in [4] is used in this work to describe and analyze DES.

Let $\Sigma$ denote a finite alphabet of events and $\Sigma^*$ the Kleene closure of all finite strings of events including the empty string $\epsilon$. The prefix closure $\mathcal{T}$ of a language $L \subseteq \Sigma$ contains all prefixes of strings in $L$. A language $L$ is said to be prefix-closed, if $L = \mathcal{T}(L)$ holds.

A DES is represented by a generator $G = (X, \Sigma, \delta, x_0)$ were $X$ denotes the finite set of states, $\Sigma$ the event alphabet, $x_0$ the initial state and $\delta : X \times \Sigma \rightarrow X$ the (partial) transition function$^1$. The notation $\delta(x, \sigma)!$ is used to indicate that $\delta(x, \sigma)$ is defined for the pair $(x, \sigma)$. The notation $\neg \delta(x, \sigma)!$ embodies the opposite. The transition function is extended to $\delta : X \times \Sigma^* \rightarrow X$ as in [4].

The uncontrolled behavior of a language $L(G) = \mathcal{L}(G)$ containing all strings that $G$ can generate. In general the event alphabet is partitioned into controllable and uncontrollable events: $\Sigma = \Sigma_c \cup \Sigma_{uc}$. This partition is fundamental within the context of controllability. The legal behavior of $G$ is denoted by a specification language $K \subseteq L(G)$. A specification language $K$ is said to be controllable w.r.t. $G$, iff

$$K \Sigma_{uc} \cap L(G) \subseteq K \quad (\text{controllability theorem}) \quad (1)$$

2.2. Partially-Strict Composed DES

Composed discrete-event systems are called partially-strict composed, if their components run in terms of a par-

---

1Marker states are not considered in this work.
tial lockstep. To be specific, the components are synchronized on some events and run asynchronously otherwise. Formally, an arbitrary subset $\Psi$ of all possible events is defined as the set of synchronized events $\Psi \subseteq (\Sigma_1 \cup \Sigma_2)$. Unlike the Synchronous Product mentioned in [4], $\Psi$ is not fixed to common events $\Sigma_1 \cap \Sigma_2$. Rather its definition is arbitrary and can be seen as a supplementary degree of freedom.

**Definition 1 (Structure of $G_1 \parallel_{\text{PSPC}} G_2$).**

Given two components $G_1$, $G_2$ defined in terms of generators. The structure of the partially-strict composed system is represented by a generator $G = G_1 \parallel_{\text{PSPC}} G_2 = (X_1 \times X_2, \Sigma, \delta, (x_0, x_0))$, with $\Sigma \subseteq \Sigma_1 \cup \Sigma_2$ the event alphabet and $\delta$ the transition function obtained by the Partially Strict Product Composition (PSPC) from [5]:

$$\delta((x_1, x_2), \sigma) =
\begin{cases}
\delta_1(x_1, \sigma) \times \delta_2(x_2, \sigma) & \text{if } \delta_1(x_1, \sigma) \land \delta_2(x_2, \sigma)!\\
\delta_1(x_1, \sigma) \times \{x_2\} & \text{if } \delta_1(x_1, \sigma)!\\
\{x_1\} \times \delta_2(x_2, \sigma) & \text{if } \delta_1(x_1, \sigma)! \land \sigma \notin \Psi\\
\{x_1\} \times \delta_2(x_2, \sigma) & \text{if } \delta_1(x_1, \sigma)! \land \sigma \notin \Psi\\
\emptyset & \text{otherwise}.
\end{cases}$$

In order to define the uncontrollable behavior, the reachability projection is introduced in [3] for an arbitrary generator $G_0 = (X_0, \Sigma_0, \delta_0, x_0)$ with $L_0 := L(G_0)$ and a generic alphabet $\Sigma_0 \supseteq \Sigma_0$.

**Definition 2 (Reachability projection).**

For $p \in \Sigma_0$, $s \in \Sigma_0$, $\sigma_1, \sigma_2 \in (\Sigma_0 \cup \{\epsilon\})$, $t = \sigma_1 \sigma_2 s \in \Sigma_0$, and $L_0 \subseteq \Sigma_0$, the reachability projection $RSP : \Sigma_0 \rightarrow \Sigma_0$ is defined as

$$RSP_{L_0}(t) = RSP_{L_0}(\epsilon, \sigma_1, \sigma_2 s) \quad \text{with}$$

$$RSP_{L_0}(\epsilon, \epsilon, \epsilon) = \epsilon$$

$$RSP_{L_0}(p, \sigma_1, \epsilon) = \begin{cases} p \sigma_1 & \text{if } p \sigma_1 \in L_0 \\
p & \text{otherwise} \end{cases}$$

$$RSP_{L_0}(p, \sigma_1, \sigma_2 s) = \begin{cases} RSP_{L_0}(p, \sigma_1, \sigma_2 s) & \text{if } p \sigma_1 \in L_0 \\
RSP_{L_0}(p, \sigma_1, s) & \text{otherwise}. \end{cases}$$

For each string $t \in \Sigma_0$, the reachability projection returns a string $u \in \Sigma_0$ by ignoring the events in $\Sigma_0 \setminus \Sigma_0$. Thus $RSP_{L_0}$ maps all strings $t \in \Sigma_0$ to the language of the generator $G_0$ [3].

**Definition 3 (Uncontrolled behavior $L(G_1 \parallel_{\text{PSPC}} G_2)$).**

Let $\Psi \subseteq \Sigma \subseteq (\Sigma_1 \cup \Sigma_2)$, $L(G_1) \subseteq \Sigma_1^*$, $L(G_2) \subseteq \Sigma_2^*$, $u \in \Sigma^*$ and $\alpha \in \Sigma$. Then $L(G_1 \parallel_{\text{PSPC}} G_2) = L_{\text{PSPC}}$ with $L_{\text{PSPC}}$ recursively defined as

1. $\epsilon \in L_{\text{PSPC}}$.
2. $(u \in L_{\text{PSPC}}) \land \left( \left( [RSP_{L(G_1)}(u) \alpha \in L(G_1)] \land (RSP_{L(G_2)}(u) \alpha \in L(G_2)) \right) \lor \left( [\epsilon \notin \Psi] \land \left( [RSP_{L(G_1)}(u) \alpha \in L(G_1)] \ight) \lor \left( RSP_{L(G_2)}(u) \alpha \in L(G_2) \right) \right) \right) \iff (u \alpha \in L_{\text{PSPC}})$.

The symbol $\oplus$ denotes the Boolean exclusive OR operator. It is proven in [3] that $L_{\text{PSPC}} = L(G_1 \parallel_{\text{PSPC}} G_2)$ holds. The following useful property states that the language both components have in common is always included in $L_{\text{PSPC}}$.

**Property 1 ([6]).** For $s \in \Sigma^*$ the following holds:

$$(s \in L(G_1)) \land (s \in L(G_2)) \Rightarrow (s \in L_{\text{PSPC}}). \quad (2)$$

3. **Problem Statement**

Inspecting Eqn. (1) clearly points out the lack of the existing controllability analysis method. $K$ and $L(G)$ are monolithic representations. Hence for a given structured model, composition has to be done a-priori risking an exponential increase of the state space. Recapitulating these considerations implies to solve the following problem by compositional analysis:

**Problem 1.** Given two local specifications $K_1 \subseteq L(G_1)$, $K_2 \subseteq L(G_2)$, locally controllable w.r.t. their respective components $G_1$, $G_2$:

$$K_1 \Sigma_{\text{uc}1} \cap L(G_1) \subseteq K_1, \quad K_2 \Sigma_{\text{uc}2} \cap L(G_2) \subseteq K_2. \quad (3)$$

Decide if their intersection $K = K_1 \cap K_2$ is globally controllable w.r.t. the partially-strict composed DES $G = G_1 \parallel_{\text{PSPC}} G_2$, without computing $G$ explicitly.

$K = K_1 \cap K_2$ is considered as the legal behavior of $G = G_1 \parallel_{\text{PSPC}} G_2$ in Probl. 1. Intersecting two different local specifications means that when considered together, the components may only perform tasks they have in common.

4. **Compositional Controllability Analysis**

Initially, all possibilities for a string to be in $L_{\text{PSPC}}$ respecting Eqn. (1) are presented. They can be verified by inspection of Def. 3. Necessary and sufficient conditions to solve Probl. 1 in a compositional manner are first listed, and then proven. Two abbreviations are introduced for the sake of clarity, namely

$$RSP_{L(G_1)}(s) \sigma := \alpha', \quad (5)$$

$$RSP_{L(G_2)}(s) \sigma := \beta'. \quad (6)$$

For all $s \in K$, $\sigma \in \Sigma_{\text{uc}}$ and $s \sigma \in L_{\text{PSPC}}$, $s \sigma \in K$ is desired, thus $s \sigma$ must respect Eqn. (1). The following cases for $s \sigma \in L_{\text{PSPC}}$ must be considered:

1. $(s \in L_{\text{PSPC}}) \land (\alpha' \in L(G_1)) \land (\beta' \in L(G_2))$: If a string $s$ is element of $L_{\text{PSPC}}$, $\sigma$ is a synchronized event and the images $RSP_{L(G_i)}(s)$ of $s$, concatenated with $\sigma$, preserve in both respective languages $L(G_i)$ ($i = 1, 2$), than $s \sigma \in L_{\text{PSPC}}$ is obtained.
2. \((s \in L_{\text{PSPC}}) \land (\sigma \notin L(G_1)) \land (\beta \notin L(G_2)) \land (\sigma \notin \Psi)\): Similar to Case 1, but \(\sigma\) is an asynchronous event and solely the image \(RSP_{L(G_1)}(s)\) of \(s\), concatenated with \(\sigma\), preserves in the respective language \(L(G_1)\).

3. \((s \in L_{\text{PSPC}}) \land (\sigma \notin L(G_1)) \land (\beta \in L(G_2)) \land (\sigma \notin \Psi)\): Similar to Case 2 with inverted roles for \(L(G_1), L(G_2)\).

Cases 2 and 3 lead to \(\sigma \notin \mathcal{K}\) and must be avoided [6]. Case 1 leads to \(s \mathcal{K}\) and \(\mathcal{K} \mathcal{K}\), thus \(s \notin \mathcal{K}\) is possible. Case 2 and 3 are avoided, if their conditions are always FALSE, thus if the following expression holds:

\[
\neg[(s \in L_{\text{PSPC}}) \land (\sigma \notin \Psi)] \land \\
\neg[(\sigma \notin L(G_1)) \land (\beta \notin L(G_2))] \land \\
\neg[(\sigma \notin L(G_1)) \land (\beta \in L(G_2))].
\]

Eqn. (7) can be reduced to the set equivalence

\[
(s \sigma \in L(G_1)) \iff (s \sigma \in L(G_2))
\]

using the rules from de Morgan and a truth table [6]. Based on Eqn. (8) a necessary condition is proposed.

**Proposition 1 (Necessary):**

For \(K_1, K_2, L(G_1)\) and \(L(G_2)\) as defined before and \(K = K_1 \cap K_2\), the following expression holds:

\[
\overline{L_{\text{PSPC}}} \cap L \subseteq \mathcal{K}
\] \Rightarrow \\
\[
\overline{L_{\text{PSPC}}} \cap L(G_1) = \overline{L_{\text{PSPC}}} \cap L(G_2).
\]

**Proof.** By contradiction, it has to be shown that assuming both \(\overline{L_{\text{PSPC}}} \cap L \subseteq \mathcal{K}\) and \(\overline{L_{\text{PSPC}}} \cap L \neq \overline{L_{\text{PSPC}}} \cap L\) implies incoherence. \(K_1 \cap K_2\) is a tautology. Assumption \(\overline{L_{\text{PSPC}}} \cap L \subseteq \mathcal{K}\) implies that \(\forall s \in \mathcal{K}\) and

\[
\forall \sigma \in \Sigma_{uc}\; (s \sigma \notin L_{\text{PSPC}}) \Rightarrow (s \sigma \notin \mathcal{K}).
\]

However, for \(\overline{L_{\text{PSPC}}} \cap L \neq \overline{L_{\text{PSPC}}} \cap L\) to be TRUE, there must exist at least one \(s \in \mathcal{K}\) and \(\sigma \in \Sigma_{uc}\), such that \((s \sigma \in \mathcal{K}) \land (s \sigma \notin \mathcal{K})\) holds. The latter one contradicts with \((s \sigma \notin \mathcal{K})\).

Subsequently, sufficiency of Prop. 1 and 2 is proposed.

**Proposition 3 (Sufficient):**

For \(K_1, K_2, L(G_1)\) and \(L(G_2)\) as defined before and \(K = K_1 \cap K_2\), the following expression holds:

\[
\overline{L_{\text{PSPC}}} \cap L \subseteq \mathcal{K}
\] \Rightarrow \\
\[
\overline{L_{\text{PSPC}}} \cap L(G_1) = \overline{L_{\text{PSPC}}} \cap L(G_2).
\]

**Proof.** By contradiction. For \(\overline{L_{\text{PSPC}}} \cap L \neq \mathcal{K}\) to be TRUE there must exist at least one \(s \sigma \in L_{\text{PSPC}}\) with \(s \in \mathcal{K}\) and \(\sigma \in \Sigma_{uc}\) such that \(s \sigma \notin \mathcal{K}\). All possible cases for \(s \sigma \in L_{\text{PSPC}}\) must be considered (cf. Def. 3):

i. \((s \sigma \in L(G_1)) \land (s \sigma \notin L(G_2))\): This case covers all strings \(s \sigma\) with \(\sigma\) an synchronized event. By Def. 2 \((s \sigma \in L(G_1)) \Rightarrow (s \sigma \sigma \in \mathcal{K})\) holds and, regarding the considered case, \((s \sigma \in L(G_1))\) holds likewise. The latter implies that \((\sigma \in \Sigma_{uc})\). Only uncontrollable events are taken under consideration, thus \(s \sigma \in \Sigma_{uc}\). Jointly it follows from Eqn. (3) that \((s \sigma \in \mathcal{K_1})\), hence the pair \((K_1, G_1)\) is locally controllable. \((s \sigma \in \overline{L_{\text{PSPC}}}K)\) follows in analogy. However, assuming \(\overline{L_{\text{PSPC}}} \cap L = \overline{L_{\text{PSPC}}} \cap \mathcal{K}_1 \cap \mathcal{K}_2\) implies that \((s \sigma \in \mathcal{K}_1) \land (s \sigma \in \mathcal{K}_2)\) leads to \((s \sigma \in \mathcal{K})\) which contradicts with \((s \sigma \notin \mathcal{K})\).

ii. \((s \sigma \in L(G_2)) \land (s \sigma \sigma \sigma \in L(G_1))\): This case covers all strings \(s \sigma\) with \(\sigma\) an asynchronous event, which can only be generated by component \(G_1\). Similarly to the previous case, the following chain of implications holds, respecting Def. 2 and the considered case: \((s \sigma \sigma \in L(G_1))\) and \((s \sigma \sigma \sigma \in L(G_2))\). However, assuming \(\overline{L_{\text{PSPC}}} \cap L \neq \overline{L_{\text{PSPC}}} \cap L\) implies that \((s \sigma \in \mathcal{K} \land (s \sigma \notin \mathcal{K} \land (s \sigma \in L(G_1)) \Rightarrow (s \sigma \sigma \sigma \in L(G_2))\) which contradicts with \((s \sigma \notin L(G_2))\).
iii. \((s \in L_{\text{PSPC}}) \land (\sigma \notin L(G_1)) \land (\beta \notin L(G_2)) \land (\sigma \notin \Psi)\): This case covers all strings \(\sigma, \beta\), with \(\sigma\) an asynchronous event, which can only be generated by component \(G_2\). In analogy to Case ii \((s \in L(G_1))\) holds. However, assuming \([\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_1)] = [\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_2)]\), implies that \((s \in K \land (\sigma \in \Sigma_{\text{uc}} \setminus \Psi)) \land (\sigma \in L(G_2))\) leads to \((s \in L(G_1))\) which contradicts with \((s \notin L(G_1))\).

The following theorem is presented, ensuring the preservation of controllability under PSPC. Note that the explicit computation of \(G = G_1 \parallel_{\text{PSPC}} G_2\) is not necessary.

\[\text{Theorem 1 (Controllability of } K \text{ w.r.t. } G)\]

\[\text{Let } \Sigma_{\text{uc}} \subseteq (\Sigma_{\text{uc1}} \cup \Sigma_{\text{uc2}}), K_i \subseteq L(G_i) = L(G) \text{ and } K_i \text{ controllable w.r.t. } G_i, i = 1, 2. K = K_1 \cap K_2 \text{ is controllable w.r.t. } G = G_1 \parallel_{\text{PSPC}} G_2 \text{ iff}\]

1. \([\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_1)] = [\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_2)],\]
2. \([\mathcal{K}_{\Sigma_{\text{uc}}} \cap K] = [\mathcal{K}_{\Sigma_{\text{uc}}} \cap K_1] \cap [\mathcal{K}_{\Sigma_{\text{uc}}} \cap K_2].\]

\[\text{Proof.} \text{ Necessity has been proven in the context of Prop. 1 and 2, sufficiency within Prop. 3.}\]

Cond. 1 of Theo. 1 only holds, if no string \(s \in [\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_1)] \cap L(G_2)\) exists, which is exclusively included in either \(L(G_1)\) or \(L(G_2)\). This leads to the following corollary:

\[\text{Corollary 1:} \text{ Cond. 1 of Theo. 1 only has a solution, if}\]

\[\left(\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_1)\right) \cap L(G_2) = \emptyset \land \left(\mathcal{K}(\Sigma_{\text{uc}} \setminus \Psi) \cap L(G_2)\right) \cap L(G_1) = \emptyset.\]

5. Illustrative Example

Theo. 1 is to be used for compositional controllability analysis of a partially-strict composed DES. Consider \(G_1, G_2\) with \(\Sigma_{\text{uc}} = \Sigma_{\text{uc1}} = \Sigma_{\text{uc2}} = \{\beta\}\), event sets, state sets and transition structures as in Fig. 1.

\[\text{Figure 1: Components}\]

Let \(\alpha, \beta\) be synchronized events and \(K_1 = \{\alpha \beta, \alpha \gamma\}, K_2 = \{\alpha \gamma, \alpha \beta \mu\}\) locally controllable w.r.t. their respective component \(G_1, G_2\). However, \(K = K_1 \cap K_2 = \{\alpha \gamma\}\) is not controllable w.r.t. the partially-strict composed DES \(G = G_1 \parallel_{\text{PSPC}} G_2\). For \(\Psi = \{\alpha, \beta\}\) and \(\Sigma_{\text{uc}} = \{\beta\}, \Sigma_{\text{uc}} \setminus \Psi = \emptyset\) holds, thus for Cond. 1 of Theo. 1 \((\{\epsilon, \alpha, \alpha \gamma\} \cap L(G_1)) \cap L(G_2) = \emptyset\) and \((\{\epsilon, \alpha, \alpha \gamma\} \cap L(G_2)) \cap L(G_1) = \emptyset\) holds using Cor. 1. Cond. 2 of Theo. 1 is not fulfilled, hence \(\mathcal{K}_{\Sigma_{\text{uc}}} \cap \mathcal{R} = \{\epsilon, \alpha, \alpha \gamma\} \beta \cap \{\epsilon, \alpha, \alpha \gamma\} = \emptyset\) and \(\mathcal{K}_{\Sigma_{\text{uc}}} \cap \mathcal{R}_1 = \{\epsilon, \alpha, \alpha \gamma\} \beta \cap \{\epsilon, \alpha, \alpha \beta, \alpha \gamma\} \cap \{\epsilon, \alpha, \alpha \beta, \alpha \gamma, \alpha \beta \mu\} = \{\alpha \beta\} \neq \emptyset\) hold. To cross-check this result, the composition is computed explicitly within the context of partially-strict composed DES (Fig. 2).

\[\text{Figure 2: Partially-strict composed DES } G = G_1 \parallel_{\text{PSPC}} G_2\]

By inspection of Fig. 2 it is verified, that \(K = \{\alpha \gamma\}\) is not controllable w.r.t. \(G = G_1 \parallel_{\text{PSPC}} G_2\). No supervisor exists which may prevent \(\beta\) in state \((2, b)\) of \(G\) from occurring.

6. Conclusions

In this paper the compositional controllability analysis approach has been extended to partially-strict composed DES to avoid the composition in the case of a partial lock-step of components. Therefor necessary and sufficient conditions were presented to ensure the conservation of controllability under partially strict product composition.

References


A New Algorithm for Diagnosability Analysis of a Class of Hierarchical Finite State Machines

Andrea Paoli† and Stéphane Lafortune‡

†CASY - DEIS, University of Bologna
Viale Risorgimento, 2 - 40136 Bologna, ITALY, Email: apaoli@deis.unibo.it

‡Department of Electrical Engineering and Computer Science, The University of Michigan
1301 Beal Avenue, MI 48109-2122, USA, Email: stephane@eecs.umich.edu

Abstract—This paper is concerned with the Fault Detection and Isolation (FDI) problem for systems modeled as Hierarchical Finite State Machines (HFSMs) using the so-called “Diagnoser Approach” introduced in prior works. A new definition of diagnosability called $L_1$-diagnosability has been studied in recent works; this property captures the possibility of detecting a fault using only high-level observations of the behavior of an HFSM. This paper deals with non-$L_1$-diagnosable HFSMs. It is shown how to study their diagnosability properties without expanding the entire HFSM model into its corresponding “flat” automaton.

1. Introduction

Many methodologies have been developed to solve the the Fault Detection and Isolation (FDI) problem for systems modeled as discrete event systems (DES). This paper is cast in the framework of the so-called ”Diagnoser Approach” to fault diagnosis of DES introduced in [1]. Consider a system modeled by an automaton denoted by $G$ that incorporates “normal” as well as “faulty” behavior for a given set of faults modeled as unobservable events. Diagnosis is the process of detecting on-line occurrences of these faults using model-based inferencing driven by the observed event sequence. In the Diagnoser Approach, this is achieved by the use of a special type of automaton, called diagnoser, which is built from the system model. In most applications, the entire system model $G$ consists of $n$ subsystems operating concurrently and coupled by parallel composition; in this case, the number of states in $G$ grows exponentially with $n$ in the worst case. This can lead to computational difficulties when using the Diagnoser Approach.

One of the methods that have been developed to mitigate the state explosion problem when modeling a DES composed of interacting components is the statecharts modeling framework of [2] and [3]. Statecharts are discrete event models embedding the idea of hierarchy (states as well as events can be seen at different levels of depth) and orthogonality (different parts of the system work in parallel). In [4] a simplified version of statecharts called Hierarchical Finite State Machines (HFSMs) is considered for solving a class of control problems under full observation. The diagnosis of systems modeled by HSFM has recently been considered in [5], leading to a semi-modular diagnosis system that exploits the structure of HFSMs as well as modularity present in the system. In [6] the class of HFSM models under investigation has been formalized and a new diagnosability definition called $L_1$-diagnosability has been introduced. This notion captures the possibility of detecting a fault using only high level observations of the behavior of an HFSM. Sufficient conditions for $L_1$-diagnosability have been presented that exploit the rich structure of HFSMs. These conditions are implementable as they are based on building extended diagnosers of projected versions of the HFSM using mostly high-level events.

In this paper the attention is focused on systems that are not $L_1$-diagnosable. The objective is to study their diagnosability properties without expanding the entire HFSM into its corresponding “flat” automaton. An algorithm that focuses on expanding only indeterminate cycles in the extended diagnoser is presented and illustrated with an example. Due to space limitations, all proofs are omitted in our presentation; these proofs, as well as several other results, can be found in [7]. Our notation is generally consistent with that in [1] and [8].

2. The notion of $L_1$-diagnosability

We assume some familiarity on the part of the reader with the class of DES denoted as HFSMs and described in detail in [6] and we only recall here some notation and results from that work.

We define a two-level HFSM to be a 6-tuple of the following form

$$G_{H} = (X; X_S; E; \delta_1; G(X_S); x_0) .$$

The set $X$ is the entire state space. The set $X_S \subseteq X$ is the set of AND super-states indexed over the set $N_{|S|} = [1 \ldots |S|]$; $x_0$ is the initial state. Super-state $x_{i}$ is composed of $n_i$ orthogonal components operating in parallel; we model the $j$-th orthogonal component of super-state $x_i$ as the automa-
ton $G_{i,j} = \{Y_{i,j}, E_{i,j}, \delta_{i,j}\}$. We denote the set of orthogonal components of $x_i$ by $G(x_i)$; moreover, $G(x_i)$ is the set of all the orthogonal components in $G_{iH}$. The set $E$ is the event set; it is composed of the following events. First, the event set of level 2 is denoted as $E_{L_2} = \bigcup_{j \in N_2} \bigcup_{e \in E_{i,j}} E_{i,j}$. Next, the event set of level 1 is denoted by $E_{L_1}$ and defined as $E \setminus E_{L_2}$. Event sets are partitioned into sets of observable and unobservable events. Namely $E_{i,j} = E_{i,j}^o \cup E_{i,j}^{no}$. We identify a subset $E_f$ of unobservable events to represent the set of fault events that can occur in the system; fault events in $E_f$ can be level 1 or level 2 events. The transition structure of $G_{iH}$ is specified entirely by the transition function at level 1, $\delta_{i,j}$, together with the set of $\delta_{i,j}$'s. The HFSMs considered in this work are assumed to satisfy Assumptions A1-A5 stated in [6] and not repeated here due to space limitations.

In [6] the notion of diagnosability introduced in [1] is adapted to HFSMs in the following manner: the objective is to investigate the possibility of detecting a fault by exploiting the structure of the HFSM. In particular, it is desired to avoid using the “flat automaton” representation of $G_{iH}$, which suffers from the state explosion problem. For this purpose, the new notion of diagnosability called L1-diagnosability is introduced.

Definition 1 L1-diagnosability  Given a fault event $f \in E_f$, we say that fault $f$ is level 1 diagnosable (L1-diagnosable) in HFSM $G_{iH}$ if for all $s \in L(G_{iH})$ that ends with event $f$, $\exists \omega \in N$ such that $\forall t \in L(G_{iH})/s$, $\exists P_{E_{i,j}}(t)$ such that $\omega \in P_{E_{i,j}}^{-1}\left[P_{E_{i,j}}(t)\right] \cap L(G_{iH}) \Rightarrow f = \omega \quad (2)$

Proposition 1 Given an HFSM $G_{iH}$, if $\forall f \in E_f$ $f$ is L1-diagnosable, then $L(G_{iH})$ is diagnosable in the sense of [1] for those faulty traces that are followed by suffixes that are live in terms of level 1 events.

In [6] a procedure to project $G_{iH}$ at level 1 is defined. The result of this procedure is a (possibly nondeterministic) automaton $G_{iH}^{L1}$ such that $L\left(G_{iH}^{L1}\right) \supseteq P_{E_{i,j}} \cdot L(G_{iH})$. A modified version of L1 projection is also considered in order to lift to level 1 information about an occurrence of level 2 fault $f_2$ in orthogonal component $G_{i,j}$ of super-state $x_i$; the resulting automaton, denoted by $G_{i,j}(f_2,a,l)$, is such that $L\left[G_{i,j}(f_2,a,l)\right] \supseteq P_{E_{i,j}} \cdot L(G_{iH})$, where $P_{E_{i,j}} \cdot L(G_{iH})$ denotes that only occurrences of $f_2$ in $G_{i,j}$ are not erased by the projection operation. Sufficient conditions to prove L1-diagnosability based on the extended diagnoser of automaton $G_{iH}^{L1}$ or $G_{i,j}(f_2,a,l)$ are presented in [6].

3. Dealing with systems that are not L1-diagnosable

Consider faults that cannot be proved to be L1-diagnosable in $G_{iH}$ using the sufficient conditions of Propositions 4 and 6 in [6]. This means that in the language of the $L_1$-projection considered in [6] (for level 1 or level 2 faults), there exist two arbitrarily long traces of events in $E_{L_1}$ (possibly with $u(f_2)$) that have the same projection with respect to $E_{L_1}^f$, but where only one of them contains the fault. These traces produce in the extended diagnoser an indeterminate cycle. Clearly, it may still be possible to diagnose such faults by using level 2 events and information about the structure of $G_{iH}$ at level 2.

We present an algorithm that is still based on the extended diagnoser of the $L_1$-projection of $G_{iH}$, but that aims to analyze the reachability or the connectivity of indeterminate cycles using level 2 events and information about the structure of $G_{iH}$ at level 2. In particular, it is sought to limit the amount of information that needs to be lifted from level 2 to level 1. First, a general algorithm to lift information from level 2 to level 1 in order to “zoom-in” a state of the extended diagnoser of the $L_1$-projection of $G_{iH}$ is presented. Then, sufficient conditions for diagnosability of the HFSM are stated and illustrated in an example.

Consider an HFSM $G_{iH}$ and its $L_1$-projection $G_{iH}^{L1}$ or $G_{iH}^{L1}(f_2,a,l)$ in the case of inclusion of a level 2 fault. To keep the notation simpler, we write $G_{iH}^{L1}$ hereafter for both cases. $G_{iH}^{L1}$ is the extended diagnoser of $G_{iH}^{L1}$. Suppose that a given state of the extended diagnoser, denoted by $\omega_{i,j}$, contains in some of its components super-state $x_i \in X_S$. Our aim is to “lift” information from level 2 in order to zoom in this state and in particular account for the structure of a relevant orthogonal component of $x_i$, denoted by $G_{i,j}$. For this purpose, we define a lifting algorithm. The lifting algorithm for extended diagnoser state $\omega_{i,j}$ entered with event $\text{entry} \in E_{L_1}^f$, with respect to a given super-state $x_i \in X_S$ and a given orthogonal component $G_{i,j}$, is described by the following steps.

Lifting algorithm for orthogonal component $G_{i,j}$ of $x_i \in X_S$ at state $\omega_{i,j}$

Input Extended diagnoser state $\omega_{i,j}$; entry event $\text{entry}$; super-state $x_i$ in $\omega_{i,j}$; orthogonal component $G_{i,j}$ of $x_i$.

Step 1 Generate the first refined extended diagnoser state by proceeding as follows for each component in $\omega_{i,j}$ entered by entry.

1.1 Substitute the given $x_i$ with the corresponding state in $Y_{i,j}$ entered by entry $\text{entry}$ for each occurrence of $x_i \in \omega_{i,j}$.

1.2 Attach a label to all states in $Y_{i,j}$ added in Step 1.1 in the usual way.

1.3 Append the names of the states of the other components $G_{i,k}$ ($k \notin N_i$, $k \neq j$) that are entered in $x_i$ by event $\text{entry}$ for each component altered in Step 1.1.

Step 2 Starting with the extended diagnoser state produced at Step 1, proceed by following the standard extended diagnoser construction procedure for automaton $G_{i,j}$ with observable events in $E_{L_1}^f$ and $E_{L_1}^o$. 

420
2.1 If the feasible event is in $E^o_{i,j}$, the states of the other orthogonal components are propagated unchanged in the various elements of the new extended diagnoser state.

2.2 If the feasible event (denoted by exit) is in $E^o_{i,j}$, proceed as described in Fig. 1 for each state element $y_{i,j} \in Y_{i,j}$ where the event exit is defined. In particular look at the corresponding exit states in $G_{i,k}$ ($k \in N_i, k \neq j$) where event exit comes from and determine if each exit state is reachable from the entry state that has been propagated with $y_{i,j}$ in the extended diagnoser component.

2.2.1 If yes, then generate the corresponding component in the next extended diagnoser state.

2.2.2 If no, then this exit event is not feasible from the extended diagnoser state component that contains $y_{i,j}$.

2.3 Repeat Steps 2.1 and 2.2 until all observable transitions within $G_{i,j}$ have been exhaustively explored and $G_{i,j}$ has been exited.

Figure 1 gives an example of Step 2.2 in the previous algorithm. Suppose that we entered $x_i$ with event entry in $y^1_{i,1}$, $y^2_{i,2}$ and $y^3_{i,3}$. We are interested in lifting information from orthogonal component $G_{1,1}$. Suppose that both $y^3_{i,1}$ and $y^1_{i,1}$ appear in the state of the extended diagnoser under consideration. Event exit $\in E^o_{i,j}$ is defined for both states. If we are in $y^3_{i,1}$ and we observe exit, then this means that the other components are either in $y^2_{i,2}$ and $y^3_{i,3}$ or in $y^1_{i,2}$ and $y^2_{i,3}$. In the first case the next state will be $x_1$. In the second case the next state will be $x_2$. Similarly if the current state in $G_{1,1}$ is $y^1_{i,1}$ and we observe exit, then the other components are in $y^2_{i,2}$ and $y^1_{i,3}$ and the next state will be $x_3$. However, since there is no connectivity between exit state $y^3_{i,3}$ and entry state $y^3_{i,3}$, then the observed transition cannot be out of state $y^3_{i,1}$ in $G_{1,1}$. Hence we were in $y^3_{i,1}$ and the new possible states are $x_1$ and $x_2$.

The above algorithm can be applied as described to refine an entire $L_1$-extended-diagnoser starting from the initial extended diagnoser state and by considering all orthogonal components. Given a specific super-state $x_i \in X_S$ that is being zoomed in, since the orthogonal components of $x_i$ have no common events (recall Assumption A1 in [6]), then the lifting algorithm can be used by treating each orthogonal component one at a time. This results in significant computational savings over building the extended diagnoser of the flat automaton as it avoids the state space explosion due to interleaving when doing the shuffle of the orthogonal components within super-states. Moreover, it allows for parallelization of the lifting algorithm. Overall, if a full lifting of all super-states and all their orthogonal components is to be performed (worst case), then the complexity of the lifting is proportional to $\prod_{i \in N_i} |N_i|$, namely all possible choices of orthogonal components for all super-states (all the lifting could be performed in parallel). On the other hand, building the extended diagnoser of the flat automaton would require expanding each super-state $x_i \in X_S$ into as many as $\prod_{i \in N_i} |N_i|$ states (worst case) before building the extended diagnoser; this is avoided by the above-described lifting algorithm.

It is possible to use the lifting algorithm to zoom in some state of interest in the $L_1$-extended-diagnoser in order to lift information from level 2 to level 1 and possibly break some connectivity in that extended diagnoser so that an indeterminate cycle may no longer be reachable or feasible.

Proposition 2 Consider fault $f_i$ at level 1 of HFSM $G_H$. Fault $f_i$ is diagnosable in the HFSM $G_H$ if it is possible to construct an extended diagnoser, built from the $L_1$-projection of $G_H$ and refined as necessary according to the lifting algorithm, that does not contain any $f_i$-indeterminate cycles.

The same result holds if fault $f_i$ is at level 2 in $G_H$, albeit in that case only occurrences of $f_i$ in orthogonal component $G_{a,i}$ are guaranteed to be diagnosable.

Example 1 Consider the HFSM $G_H$ in Fig. 2. The system can be affected by a level 1 fault $f_i$. The $L_1$-projection of $G_H$ as well as its extended diagnoser are shown in Fig. 3. Since the extended diagnoser of the $L_1$-projection of $G_H$ contains an indeterminate cycle, Proposition 4 in [6] does not allow us to conclude about the $L_1$-diagnosability of $f_i$ in $G_H$. 

Figure 1: Application of Step 2.2 of the lifting algorithm.
4. Conclusions

In this work we presented new results on diagnosability analysis of a special class of HFSMs. Starting from the notion of $L_1$-diagnosability, we have addressed the case of faults that cannot be proved to be $L_1$-diagnosable using the results in [6]. In order to avoid the complexity of working with the flat automaton, we have presented an algorithm to refine the extended diagnoser built from the higher level projection of the HFSM by lifting information from the lower level. This lifting algorithm led to Proposition 2 which presents a sufficient condition to test diagnosability using the refined version of the diagnoser.

References


A numerical verification method for ODEs based on Nakao’s theory

Nobito Yamamoto
†
†Department of Computer Science
The University of Electro-Communications
1-5-1 Chofugaoka Chofu, Tokyo, 182-8585, Japan
Email: yamamoto@im.uec.ac.jp

Abstract—Numerical verification methods on existence or uniqueness of solutions to PDEs, so called Nakao’s theory[1][2], have been developed by M.T.Nakao in Kyushu University, Japan, and his group including the author. They are based on fixed point theorems and error estimation of approximate solutions which are mainly computed by FEM.

On the other hand, there is a standard method for numerical verification of ODEs named the Lohner method[3]. It uses Taylor expansion and the error estimation of truncation errors. Wrapping effect might be reduced by using QR factorization. However, the method validates the values of the solution only just on the nodal points and tells nothing about the error which occurs in the intervals between the nodes.

We present a numerical verification method based on stepwise application of Nakao’s theory which guarantees the error bounds not only on the nodes but also between the nodes. QR factorization is used to reduce the wrapping effect. We intend to develop this method to apply to the evolution equations in the future.

1. Nakao’s theory

Nakao’s theory consists of numerical verification methods to prove existence or uniqueness of solutions to PDEs. Outline of the numerical verification methods can be described as follows [1][2].

1. Compute an approximate solution \( \tilde{u} \) to the equation by an appropriate method.

2. Derive an equation for the error \( w \) of the approximation \( \tilde{u} \).

3. Transform the error equation into a fixed point form. Consider a function space \( X \) which contains solutions to the error equation.

4. Devide \( w \) into \( w_x \in S_N \) and \( w_\perp \in S_\perp \), where \( S_N \subset X \) is a finite dimensional subspace which approximates \( X \) in some sense and \( S_\perp \) is the complementary space of \( S_N \). Transform the fixed point equation into a system of two kinds of equations, the equations for \( w_x \) and the equations for \( w_\perp \).

5. Transform the equations for \( w_x \) to the quasi Newton form.

6. Construct another fixed point form using the quasi Newton form and the equations for \( w_\perp \).

7. Apply a suitable fixed point theorem to the above fixed point form and derive sufficient conditions for \( W \subset X \) in order that there exists a solution or the unique solution within the candidate set \( W \).

8. Define \( W := W_N + W_\perp \) where \( W_N \subset S_N \) and \( W_\perp \subset S_\perp \). Separate the sufficient conditions into 2 parts, conditions for \( W_N \) and conditions for \( W_\perp \). Moreover, the conditions for \( W_\perp \) are modified into computable form using error analysis of the approximate subspace \( S_N \) to \( X \).

9. Construct the candidates for \( W_N \) and \( W_\perp \) by some iterative method, and verify the sufficient conditions using self-validated numerics.

2. Problem

We treat an ordinary differential equation with initial value conditions as follows.

\[
\begin{align*}
\frac{du}{dt} &= f(t,u), \quad t \in [0,T] \\
u(0) &= u_0
\end{align*}
\]

Here \( u \) is a vector function in \( \mathbb{R}^m \) and \( f(t,u) \) is an operator from \( \mathbb{R} \times \mathbb{R}^m \) to \( \mathbb{R}^m \), which is continuous with respect to \( t \) and Fréchet differentiable with respect to \( u \).

Take nodal points between \([0,T]\) as \( t_0 = 0 < t_1 < \cdots < t_n = T \) and compute an approximate solution \( \tilde{u} \). From theoretical point of view, the computation can be done by an arbitrary method, but there is an appropriate method described in the section \( 5 \).

Put \( \tilde{u}_j := \tilde{u}(t_j) \). Let \( w \) be the error as \( w(t) := u(t) - \tilde{u}(t) \) and \( w_j := w(t_j) \). We consider an error equation from (1), which is restricted within \([t_{j-1}, t_j]\) for fixed \( j \in \{1,2,\ldots,n\} \) as follows.

\[
w(t) = w_{j-1} + \int_{t_{j-1}}^{t_j} f(\tau, \tilde{u} + w)d\tau - (\tilde{u}(t) - \tilde{u}_{j-1}).
\]
Note that this is a fixed point equation. The solution \( w \) is to be found in a space \( X := (C(t_{j-1}, t_j))^m \).

Now we take \( S_N \subset X \). The subspace \( S_N \) is the set of linear polynomials on \( [t_{j-1}, t_j] \), which is of \( 2 \times m \) dimension, because every function in \( S_N \) is specified by its values at the end points \( t_{j-1} \) and \( t_j \). In order to make sense that \( S_N \) approximates \( X \), we introduce \( P_N : X \to S_N \) as an operator that gives the first interpolation polynomial of \( x \in X \) which interpolates \( x \) on \( [t_{j-1}, t_j] \). This is a projection from \( X \) to \( S_N \), and we can write \( S_N = P_N X \).

Take \( w_N := P_N w \) and \( w_\perp := (I - P_N)w \), and separate the error into \( w = w_N + w_\perp \). Transform the error equation (2) to a system of two parts.

\[
w_n(t) = w_{j-1} + r(\hat{u}) + P_N \int_{t_{j-1}}^{t} \left[ f(\tau, \hat{u} + w_N + w_\perp) - f(\tau, \hat{u}) \right] d\tau,
\]
and

\[
w_\perp(t) = (I - P_N) \int_{t_{j-1}}^{t} f(\tau, \hat{u} + w_N + w_\perp) d\tau,
\]
where \( r(\hat{u}) \) is defined by

\[
r(\hat{u}) := P_N \int_{t_{j-1}}^{t} f(\tau, \hat{u}) d\tau - (\hat{u}(t) - \hat{u}_{j-1}).
\]

Now we transform (3) to the quasi Newton form. Define \( F(w_n; w_\perp) \) by

\[
F(w_n; w_\perp) = w_{j-1} + r(\hat{u}) + P_N \int_{t_{j-1}}^{t} \left[ f(\tau, \hat{u} + w_N + w_\perp) - f(\tau, \hat{u}) \right] d\tau,
\]
where \( w_\perp \) is fixed and considered as a parameter. Note that the Fréchet derivative of \( F \) with respect to \( w_n \) is given by

\[
F'(w_n; w_\perp)w = P_N \int_{t_{j-1}}^{t} \frac{\partial}{\partial u} f(\tau, \hat{u} + w_N + w_\perp) \cdot v(\tau) d\tau.
\]

Take \( F^* := F(0; 0) \) and derive the quasi Newton form of (3) as follows.

\[
w_n = w_n - (I - F^*)^{-1} \left[ w_N - F(w_n; w_\perp) \right]
\]

\[
= (I - F^*)^{-1} \left[ F(w_n; w_\perp) - F(w_n) \right]
\]

\[
= (I - F^*)^{-1} \left[ w_{j-1} + r(\hat{u}) \right] + P_N \int_{t_{j-1}}^{t} \left[ f(\hat{u} + w_N + w_\perp) - f(\hat{u}) - \frac{\partial f}{\partial u} \bigg\vert_{u=\hat{u}} w_N \right] d\tau.
\]

Let \( N(w_n, w_\perp) \) denotes the right-hand side. Combining (8) and (4), we define

\[
T(w) := N(P_N w, (I - P_N)w) + (I - P_N) \int_{t_{j-1}}^{t} f(\hat{u} + w_N) d\tau,
\]
and obtain another fixed point form

\[
w = T(w)
\]
which is equivalent to (2).

### 3. Sufficient Conditions

Since the operator \( T \) is compact on \( X \), we apply Schauder's fixed point theorem to the fixed point form (10). Then if we find a bounded convex closed set \( W \subset X \) such that

\[
T(W) \subset W
\]
holds, it is proved that there exists a solution \( w \) to the equation (2) within \( W \). Hereafter we deal with just such a set \( W \) that has the following form.

\[
W := W_N + W_\perp = [w_N + w_\perp | w_N \in W_N, w_\perp \in W_\perp]
\]
for some \( W_N \subset S_N \) and \( W_\perp \subset S_\perp \). And we call \( W \) a candidate set. From (11)

\[
T(W) \subset W \Leftrightarrow \left\{ \begin{array}{l}
P_N T(W) \subset W_N \\\\\\\\ (I - P_N) T(W) \subset W_\perp,
\end{array} \right.
\]
and then

\[
T(W) \subset W \Leftrightarrow \left\{ \begin{array}{l}
N(W_N, W_\perp) \subset W_N \\\\\\\\ (I - P_N) \int_{t_{j-1}}^{t} f(\hat{u} + w_N + w_\perp) d\tau \subset W_\perp.
\end{array} \right.
\]

holds. Here

\[
N(W_N, W_\perp) = (I - F^*)^{-1} \left[ w_{j-1} + r(\hat{u}) \right] + P_N \int_{t_{j-1}}^{t} \left[ f(\hat{u} + w_N + w_\perp) - f(\hat{u}) - \frac{\partial f}{\partial u} \bigg\vert_{u=\hat{u}} W_N \right] d\tau.
\]

Though we omit the detail, sufficient conditions to the first expression of (13) are described by

\[
(I - A)^{-1} \left[ w_{j-1} + r_j + \int_{t_{j-1}}^{t} \left[ f(\hat{u} + \tilde{w}_n) - f(\hat{u}) \right] d\tau \right] + (A' - A)w_n(t_j) + \int_{t_{j-1}}^{t} \frac{\partial f}{\partial u} \bigg\vert_{u=\hat{u}+\tilde{w}_n+w_\perp} w_\perp d\tau \in W_N.
\]

for arbitrary \( w_n \in W_N \) and \( w_\perp \in W_\perp \), where

\[
\tilde{w}_n(t) := \frac{t_j - t}{t_j - t_{j-1}} w_{j-1}, \quad t \in [t_{j-1}, t_j],
\]

\[
W_n^* := w_N - \tilde{w}_n,
\]

\[
r_j := r(\hat{u}) \bigg\vert_{t=t_j},
\]

\[
A := \int_{t_{j-1}}^{t_j} \frac{\partial f}{\partial u} \bigg\vert_{u=\hat{u}+\tilde{w}_n+w_\perp} w_\perp d\tau,
\]

\[
A' := \int_{t_{j-1}}^{t_j} \frac{\partial f}{\partial u} \bigg\vert_{u=\hat{u}+\tilde{w}_n+w_\perp} w_\perp d\tau.
\]

In order to modify the second expression of (13), we adopt the following proposition.
Proposition 1 If \( \frac{dx}{dt} \) is bounded in \((t_j, t_{j+1})\) for \( x \in X \), then
\[
\| (I - P_N) x \|_{\infty(j)} \leq \frac{1}{2} (t_j - t_{j-1}) \| \frac{dx}{dt} \|_{\infty(j)}
\] (21)
holds. Here
\[
\|v\|_{\infty(j)} = \sup_{t \in [t_{j-1}, t_j]} |v|
\] (22)
which is defined elementwise.

This proposition gives an error estimation of \( P_N x \in S_N \) to \( x \in X \).

4. How to construct the candidate set

In order to construct the candidate set, we give \( W_N \) and \( W_\perp \) as follows.
\[
W_N = W_N(w_{j-1}, \tilde{\alpha}_j)
\]
\[
:= \{ w_s \in S_N | w_s(t_{j-1}) = w_{j-1}, [B_j w_s(t_j)] \leq \tilde{\alpha}_j \}
\] (23)
and
\[
W_\perp = W_\perp(\beta_j)
\]
\[
:= \{ w_s \in S_\perp | \| w_s \|_{\infty(j)} \leq \beta_j \}
\] (24)
Here \( \tilde{\alpha}_j \) and \( \beta_j \) are constant vectors in \( \mathbb{R}^m \) which should be specified in order that the sufficient conditions hold. And \( \| \cdot \| \) implies the absolute value matrix, namely a matrix taking the absolute value of every element. The matrix \( B_j \in \mathbb{R}^{m \times m} \) is defined as described below in order to reduce the wrapping effect. The value \( w_{j-1} \) is bounded by
\[
|B_j w_{j-1}| \leq \tilde{\alpha}_j
\] (25)
through the previous step. Let \( I_m \) be the \( m \times m \) identity matrix and
\[
[-I_m, I_m] := \begin{pmatrix}
-1, 1 \\
-1, 1 \\
\vdots \\
-1, 1
\end{pmatrix}
\]
as an interval matrix. Define the matrix \( A_1 \) by
\[
A_1 := \int_{t_{j-1}}^{t_j} \frac{\partial f}{\partial u} \big|_{u = \tilde{u}} dt_j - t_{j-1} d\tau
\] (26)
and recast \( A_2 := A \). Moreover let us take some interval matrices as follows.
\[
[A_1] := \int_{t_{j-1}}^{t_j} \frac{\partial f}{\partial u} \big|_{u = \tilde{u} + \tau e_{j-1}} dt_j - t_{j-1} d\tau
\] (27)
\[
[A_2] := \int_{t_{j-1}}^{t_j} \frac{\partial f}{\partial u} \big|_{u = \tilde{u} + e_{j-1}} \big|_{\tau = \frac{t_j - t_{j-1}}{\tau}} dt_j - t_{j-1} d\tau
\] (28)
where
\[
[\alpha_{j-1}] := [-I_m, I_m] [B_{j-1}^{-1}] \tilde{\alpha}_{j-1},
\] (29)
and
\[
[A_3] := \int_{t_{j-1}}^{t_j} \frac{\partial f}{\partial u} \big|_{u = \tilde{u} + \tau e_j} dt_j
\] (30)

Then the condition (15) is satisfied if
\[
|B_j(I - A_2)^{-1}(I + A_1) B_{j-1}^{-1} [-I_m, I_m] \tilde{\alpha}_{j-1} + [B_j(I - A_2)^{-1} [-I_m, I_m] \tilde{\alpha}_j)
\]
\[
+ [B_j(I - A_2)^{-1} ([A_1] - A_1) B_{j-1}^{-1} [-I_m, I_m] \tilde{\alpha}_{j-1}]
\]
\[
+ [B_j(I - A_2)^{-1} ([A_2] - A_2) B_{j-1}^{-1} [-I_m, I_m] \tilde{\alpha}_j)
\]
\[
+ [B_j[A_3]] [-I_m, I_m] \beta_j \subset [-I_m, I_m] \tilde{\alpha}_j
\]
holds. Note that \( T[-I_m, I_m] a \subset [-I_m, I_m] T[a \) for arbitrary \( m \times m \) matrix \( T \) and \( m \) vector \( a \). Then this condition implies an inequality
\[
|B_j(I - A_2)^{-1}(I + A_1) B_{j-1}^{-1} \tilde{\alpha}_{j-1} + [B_j(I - A_2)^{-1} \cdot [r_j]
\]
\[
+ [B_j(I - A_2)^{-1} ([A_1] - A_1) B_{j-1}^{-1} \tilde{\alpha}_{j-1}]
\]
\[
+ [B_j(I - A_2)^{-1} ([A_2] - A_2) B_{j-1}^{-1} \tilde{\alpha}_j)
\]
\[
+ [B_j[A_3]] \beta_j \leq \tilde{\alpha}_j.
\] (31)
From the left-hand side of the inequality, it is better to take \( r_j \) as small as possible, which suggests us the way to compute the approximate solution. We compute \( \hat{u} \) by the Newton method in order that
\[
P_N \int_{t_{j-1}}^{t_j} f(\tau, \hat{u}) d\tau - (\hat{u} - \tilde{u}_{j-1}) = 0
\]
holds approximately. It can be seen that this approximation method is of \( O(h^2) \) where \( h \) denotes the maximum width of \( t_j - t_{j-1} \) for \( j = 1, \ldots, n \).

Now we take \( B_0 = I_m \) and specify \( B_j \) such that the matrix
\[
B_j(I - A_2)^{-1}(I + A_1) B_{j-1}^{-1}
\]
has the least effect of rotation. Therefore we compute QR factorization as
\[
(I - A_2)^{-1}(I + A_1) B_{j-1}^{-1} = Q_{j-1} R_{j-1}
\] (32)
and define
\[
B_j = Q_{j-1}^{-1}
\] (33)
This device might reduce the wrapping effect.

On the other hand, a sufficient condition for the second expression of (13) is given as follows.
\[
\frac{1}{2} (t_j - t_{j-1}) \| f(\cdot, \tilde{u} + W_N + W_\perp) - \epsilon \|_{\infty(j)} \leq \beta_j.
\] (34)
Here we used (21) on the error estimation of \( P_N \). Since the constant vector \( \epsilon \) can be taken arbitrarily, we take \( \epsilon \) as the midpoint of the interval vector \( f([t_j, t_{j-1} - \tilde{u} + W_N + W_\perp] \).

For \( j = 1, \ldots, n \), we specify \( \tilde{\alpha}_j \) and \( \beta_j \) satisfying the conditions (31) and (34) by some iterative method and \( B_j \) by (33) step by step.
5. Numerical example

In our talk, we will deal with a simple problem as follows.

\[
\begin{align*}
\frac{d^2u}{dt^2} &= -u + u^2, \quad 0 < t < T \\
\frac{du}{dt}(0) &= u_0 \\
u'(0) &= u_0' 
\end{align*}
\]  

(35)

Taking \(u_1(t) = u(t), \quad u_2(t) = u'(t),\)

the equation (35) is transformed to

\[
\begin{align*}
\frac{du}{dt} &= f(u) \\
u(0) &= u_0
\end{align*}
\]

(36)

with

\[
\begin{align*}
u(t) := \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \quad u_0 := \begin{pmatrix} u_0 \\ u_0' \end{pmatrix} \\
f(u) := \begin{pmatrix} f_1(u) \\ f_2(u) \end{pmatrix} = \begin{pmatrix} u_2 \\ -u_1 + u_2^2 \end{pmatrix}
\end{align*}
\]

We intend to show a comparison with the Lohner method and some improvement using higher order interpolations in the presentation.

6. Acknowledgment

The author appreciates Mr. M. Fukuhara, assistant professor in the University of Electro-Communications, for his useful suggestion and efforts to improve the manuscript.

References


Some computer assisted proofs on three dimensional heat convection problems

M.-N. Kim¹, M. T. Nakao¹, Y. Watanabe² and T. Nishida³

¹ Faculty of Mathematics, Kyushu University 33, Fukuoka 812-8586, Japan
² Computing and Communications Center, Kyushu University 33, Fukuoka 812-8586, Japan
³ Graduate School of Science and Engineering, Waseda University, Tokyo 169-8555, Japan

Email: ¹ {mnkim, mtnakao}@math.kyushu-u.ac.jp
² watanabe@cc.kyushu-u.ac.jp
³ tkninish@waseda.jp

Abstract—A computer assisted proof of the existence of nontrivial steady state solutions for the three dimensional Rayleigh–Bénard convection is described based on the fixed point theorem using a Newton like operator.

1. Introduction

The Rayleigh–Bénard convection is a heat convection between two infinite solid plates with hot bottom and cool top. The motion of fluid is self-sustained as soon as gravitational energy release overcomes dissipation losses, which is called Rayleigh’s mechanism by buoyancy [2]. We use the Oberbeck–Boussinesq equations as approximate equations for this problem with normalization of variables and parameters as follows:

\[
\begin{align*}
\frac{1}{P} \frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p &= \mathbf{u} - (\mathbf{G} - \mathbf{RT}) \mathbf{e}_3, \quad (1a) \\
\nabla \cdot \mathbf{u} &= 0, \quad (1b) \\
\frac{\partial T}{\partial t} + (\mathbf{u} \cdot \nabla) T &= \Delta T, \quad (1c)
\end{align*}
\]

where \( \mathbf{u} \) the velocity field, \( p \) the pressure, \( T \) the temperature of the fluid, \( \mathbf{R} \) Rayleigh number, \( P \) Prandtl number, and a parameter \( \mathbf{G} \). We use stress free boundary conditions for velocity field, \( T = 0 \) on \( x_3 = 0 \) and \( T = \pi \) on \( x_3 = \pi \). The equilibrium state is the pure heat conduction solution \( \mathbf{u} = 0, T = \pi - x_3, \) and \( p = \mathbf{G}(\pi - x_3) - \frac{1}{2}(\pi - x_3)^2 + p_0 \) for any \( R > 0 \) and \( P > 0 \). Here \( p_0 \) is a reference pressure. Using this conduction solution \( (\mathbf{u}, T, p) \) and eliminating time derivatives from (1), the steady state bifurcation equations for the perturbation \( (\mathbf{u}, \theta, p) \) to the equilibrium are given by

\[
\begin{align*}
-\Delta \mathbf{u} + \frac{1}{P} (\mathbf{u} \cdot \nabla) \mathbf{u} + \nabla p - \mathbf{R} \mathbf{e}_3 &= 0, \quad (2a) \\
\nabla \cdot \mathbf{u} &= 0, \quad (2b) \\
-\Delta \theta + (\mathbf{u} \cdot \nabla) \theta - \mathbf{u}_0 &= 0. \quad (2c)
\end{align*}
\]

Assume that all fluid motion is confined to \( \Omega = \{0 < x_3 < 2\pi/a, 0 \leq x_2 \leq 2\pi/b, 0 \leq x_1 \leq \pi \} \) for given wave numbers \( a, b > 0 \) and impose parity conditions on new boundaries as in [3] which lead periodic boundary conditions in horizontal directions [4]. From these boundary conditions, the velocity field, the perturbation of temperature and pressure can be represented by the following Fourier series [3]:

\[
\mathbf{u} = \sum_{\alpha \neq 0} \sum_{j=1}^3 \mathbf{e}_j u_\alpha^{(j)}, \quad \theta = \sum_{\alpha \neq 0} \theta_\alpha^{(j)}, \quad p = \sum_{\alpha \neq 0} p_\alpha^{(j)}, \quad (3)
\]

where \( \alpha \in \mathbb{Z}_3 \), \( \mathbb{Z}_0 \) is the set of all non–negative integers, and \( u_\alpha^{(j)}, \theta_\alpha^{(j)}, p_\alpha^{(j)} \) coefficients of \( \mathbf{u}, \theta, p \) with respect to the base functions \( \phi_\alpha^{(j)} \) defined by,

\[
\begin{align*}
\phi_\alpha^{(1)}(x) &= K_0 \sin(\alpha_1 x_1) \cos(\alpha_2 x_2) \cos(\alpha_3 x_3), \quad (4a) \\
\phi_\alpha^{(2)}(x) &= K_0 \cos(\alpha_1 x_1) \sin(\alpha_2 x_2) \cos(\alpha_3 x_3), \quad (4b) \\
\phi_\alpha^{(3)}(x) &= K_0 \sin(\alpha_1 x_1) \cos(\alpha_2 x_2) \sin(\alpha_3 x_3), \quad (4c) \\
\phi_\alpha^{(4)}(x) &= K_0 \cos(\alpha_1 x_1) \cos(\alpha_2 x_2) \cos(\alpha_3 x_3), \quad (4d)
\end{align*}
\]

where the normalization factor \( K_0 = \prod_{i=1}^3 \sqrt{2 - \delta_{0i0}} \), \( 2 \sqrt{2} K_0 \). Here \( \delta_{ij} \) is the usual Kronecker delta on indices \( i \) and \( j \). We introduce function spaces for \( i = 1, \ldots, 4 \) and \( k \geq 0 \), \( H^k(\Omega) = \left\{ \sum_{\alpha \neq 0} u_\alpha^{(j)} \right\}_{\sum_{\alpha \neq 0} u_\alpha^{(j)} A_\alpha^{(j)} < \infty} \subset H^k(\Omega) \), where \( L^2(\Omega) = H^0(\Omega) \subset L^2(\Omega) \) and for \( \alpha \in \mathbb{Z}_3^* \)

\[
A_\alpha^{(j)} = (\alpha_1^2 + (\alpha_2^2) + \alpha_3^2, \quad B_\alpha^{(j)} = (aa_1)^2 + (ba_2)^2. \quad (5)
\]

And we also introduce a divergence free function space \( V^k = \{ \mathbf{u} \in H^k(\Omega) \times H^k(\Omega) \times H^k(\Omega) | \nabla \cdot \mathbf{u} = 0 \} \) and set \( W^k = H^k(\Omega) \times H^k(\Omega) \) for \( \mathbf{v} \). And \( V^k \) has base functions:

\[
\begin{align*}
\Phi_\alpha^{(1)} &= \frac{-a_1 a_3 \phi_{\alpha_3}^{(1)}}{A_\alpha B_\alpha} \phi_\alpha^{(1)}, \frac{b_2 a_3 \phi_{\alpha_3}^{(2)}}{A_\alpha B_\alpha} \phi_\alpha^{(2)}, \frac{B_\alpha \phi_{\alpha_3}^{(3)}}{A_\alpha}, \alpha \in S^1, \quad (6a) \\
\Psi_\alpha^{(1)} &= \frac{b_2 a_3^{(1)}}{B_\alpha} \phi_{\alpha_3}^{(1)}, -\frac{a_1 a_3^{(1)}}{B_\alpha} \phi_\alpha^{(1)}, \alpha \in S^2, \quad (6b)
\end{align*}
\]

where \( S^1 = \{ \alpha \in S | \alpha_3 = 0 \}, S^2 = \{ \alpha \in S | \alpha_1 \notin S \} \) and \( S = \mathbb{Z}_0^* - \alpha_3 \mathbb{Z}_0 + \alpha_2 \mathbb{Z}_0. \) Clearly \( S = S^1 \cup S^2 \). Thus any \( \mathbf{u} \in V^k \) can be expressed as \( \mathbf{u} = \sum_{\alpha \in S} u_\alpha \Phi_\alpha^{(j)} + v_\alpha \Psi_\alpha^{(j)} \) where \( u_\alpha = 0 (\alpha \in S - S^1), v_\alpha = 0 (\alpha \in S - S^2) \). Similarly, \( \mathbf{v} \in W^k \) can be represented in the form \( \mathbf{v} = \sum_{\alpha \in \mathbb{S}} \theta_\alpha \phi_\alpha^{(j)} \), and \( T = \{ \alpha \in \mathbb{Z}_3^* | \alpha_3 = 0 \} \) is the index set for \( W^k \).

2. A priori error estimates

For a fixed positive integer \( N \), define infinite dimensional subspaces \( V_N = \{ \mathbf{u} \in V^k | u_\alpha = v_\alpha = 0, \forall \alpha \in S, |\alpha| > N \} \),
and $W_N = \left\{ \theta \in W^k \mid \theta_\alpha = 0, \forall \alpha \in T, |\alpha| > N \right\}$. Here $|\alpha| = a_1 + a_2 + a_3$. And define projections $P_N : V^k \rightarrow V_N$ and $Q_N : W^k \rightarrow W_N$ as follows[4]:

\begin{align}
(\nabla (u - P_N u), v) &= 0, \quad \forall v \in V_N; \tag{7a} \\
(\nabla (\theta - Q_N \theta), \nabla \theta) &= 0, \quad \forall \theta \in W_N. \tag{7b}
\end{align}

Due to orthogonal relations of base functions of $V^k$ and $W^k$, these projections $P_N$ and $Q_N$ are truncation operators, i.e., $P_N u = \sum_{\alpha \in T_N} \alpha u^\alpha$ and $Q_N \theta = \sum_{\alpha \in T_N} \alpha \theta^\alpha$, where $S_N = \left\{ \alpha \in S \mid |\alpha| \leq N \right\}$ and $T_N = \left\{ \alpha \in T \mid |\alpha| \leq N \right\}$. From these characterization of projections, we have

**Theorem 1** For any $u \in V^k$, $\theta \in W^k$ with $k \geq 2$, and their finite dimensional approximations $P_N u \in V_N$, $Q_N \theta \in W_N$ in (7), we have a priori error estimates:

\begin{align}
||u - P_N u||_{L_2} &\leq C_1 ||\Delta u||_{L_2}, \quad ||\theta - Q_N \theta||_{L_2} \leq C_2 ||\Delta \theta||_{L_2}; \tag{8a} \\
||u - P_N u||_{L_\infty} &\leq ||\Delta u||_{L_\infty}, \quad ||\theta - Q_N \theta||_{L_\infty} \leq ||\Delta \theta||_{L_\infty}. \tag{8b}
\end{align}

where $C_0 = a_1^2 + b_2^2 + 1/(n+1)$.

Applying Cauchy–Schwarz–Schur inequality on the estimated bound of $||u, u^\alpha + v^\alpha, v^\alpha||_{L_\infty}$, we obtain

**Lemma 2** For $u \in V^k$, $\theta \in W^k$ with $k \geq 2$, it holds that

\[ ||u||_{L_\infty} \leq C_1 ||\Delta u||_{L_2}, \quad ||\theta||_{L_\infty} \leq C_2 ||\Delta \theta||_{L_2}, \tag{9} \]

where $C_1 = \left[ \sum_{\alpha \in T_N} K_2^\alpha A_{\alpha} \right]^{1/2}$ and $C_2 = \left[ \sum_{\alpha \in T - T_N} K_2^\alpha A_{\alpha} \right]^{1/2}$.

**Corollary 3** Under the same assumptions of Theorem 1, the following a priori error estimates hold:

\[ ||u - P_N u||_{L_\infty} \leq C_3 ||\Delta u||_{L_2}, \quad ||\theta - Q_N \theta||_{L_\infty} \leq C_4 ||\Delta \theta||_{L_2}, \tag{10} \]

where $C_3 = \left[ \sum_{\alpha \in T - T_N} K_2^\alpha A_{\alpha} \right]^{1/2}$, $C_4 = \left[ \sum_{\alpha \in T - T_N} K_2^\alpha A_{\alpha} \right]^{1/2}$.

It’s easy to show that $C_1, C_2 \leq 5.45 K_0 (a_1^2 + b_2^2 + 1)$ and $C_3, C_4 = O(N^{1/4})$.

The steady state equation of (2) satisfies the following Stokes equations:

\begin{align}
-\Delta u + \nabla p &= f(u, \theta), \quad \text{\text{(11a)}} \\
\nabla u &= 0, \quad \text{\text{(11b)}} \\
-\Delta \theta &= g(u, \theta), \quad \text{\text{(11c)}}
\end{align}

where $f = -\frac{1}{\rho} (u \cdot \nabla u) + R \phi$, and $g = -(u \cdot \nabla \theta) + u_3$. If $(u, \theta) \in V^2 \times W^2$, then $(-\Delta u, -\Delta \theta) \in L^2(\Omega)^2$ and $\nabla(-\Delta u) = -\Delta(\nabla u) = 0$, since all combinations are functions of sinusoidal functions. Moreover, $(-\nabla p, -\Delta u) = (p, \nabla \Delta u) = 0$ (see [1]). Taking inner product on (11) with $-\Delta u$ and $-\Delta \theta$, we have $||\Delta u||^2 = (f, -\Delta u) \leq ||f||||\Delta u||$ and $||\Delta \theta||^2 = (g, -\Delta \theta) \leq ||g||||\Delta \theta||$. This gives us the following estimates:

\[ ||\Delta u||_{L_2} \leq C_0 ||f||_{L_2}, \quad ||\Delta \theta||_{L_2} \leq C_0 ||g||_{L_2}. \tag{12} \]

We will use the symbol $S$ for the solution operator of (11) and call it as Stokes operator.

### 3. A fixed point formulation

For $(u, \theta) \in X = V^1 \times W^1$, (2) can be written in the following variational form: $\phi(v, \theta) \in X$,

\[ (\nabla u, \nabla v) + \frac{1}{\rho} (u \cdot \nabla u, v) - (\mathcal{R}(\theta, v_3)) = 0, \tag{13a} \]

\[ (\nabla \theta, \nabla \theta) + (u \cdot \nabla \theta, \theta) - (u_3, \theta) = 0 \tag{13b} \]

which can be written in simpler form: find $\phi = (u, \theta) \in X$ such that $(\nabla \phi, \nabla \phi) = (F(\theta, \phi), (F(\phi), v))$ for $\phi = (v, \theta) \in X$. We denote an approximate solution $\phi_N \in X_N = V_N \times W_N$, i.e., $(\nabla \phi_N, \nabla \phi_N) = (F(\phi_N), \phi_N)$, for $\phi_N \in X_N$. Usually we use Newton’s method to get an approximation and from now on, we denote it as $\phi_N \in X_N$, i.e., $(\nabla \phi_N, \nabla \phi_N) \approx (F(\phi_N), \phi_N)$ for $\phi_N \in X_N$.

Let $\phi \in X$ be the solution of the equation: $(\nabla \phi, \nabla \phi) = (F(\phi_N + \phi), \phi) - (\nabla \phi_N, \nabla \phi_N) = (G(\phi), \phi)$ for $\phi \in X$, i.e., the residual equation. It can be rewritten as a fixed point equation:

\[ \phi = S \phi = K \phi. \tag{14} \]

Since $X \subset H^1(\Omega)^4$, $K$ is a compact operator in $X$. Hence by the Schauder’s fixed point theorem, if we find a nonempty, closed, bounded and convex set $U \subset X$ satisfying $K U \subset U$, then there exists a solution of (14) in $U$. The set $U$ is called a candidate set of solutions.

Define $\mathcal{P}_N : X \rightarrow X_N$ by $\mathcal{P}_N = (P_N, Q_N)$, then (7) can be simplified as: $(\nabla \phi - \mathcal{P}_N F, \nabla \phi) = 0$, for $\phi \in X$. Then (14) can be decomposed into two parts:

\[ \mathcal{P}_N \phi = \mathcal{P}_N \mathcal{K} \phi, \tag{15a} \]

\[ (I - \mathcal{P}_N) \phi = (I - \mathcal{P}_N) K \phi. \tag{15b} \]

Now assume the regularity of the operator: $\mathcal{L} = \mathcal{P}_N [I - SF(\phi_N)] \mid_{X_N} : X_N \rightarrow X_N$, i.e., $\mathcal{L}$ is one-to-one and onto. Here the Fréchet derivative $F'(\phi)$ of $F$ at $\phi$ has the form: for any $\phi \in X$, $F'(\phi) = (F'(\phi) \phi, g' (\phi) \phi)$. Here $F'(\phi) \phi = -\frac{1}{\rho} [u \cdot \nabla v + (v \cdot \nabla) u + \mathcal{R} \phi_3 + g' (\phi) \phi]$. $g' (\phi) \phi = -\frac{1}{\rho} [u \cdot \nabla \theta + (v \cdot \nabla) \theta + v_3]$ for $\phi = (u_\theta, \phi) = (v, \theta) \in X$. Note that $\mathcal{L}$ can be expressed as: $\mathcal{L} = \mathcal{P}_N [I - SF(\phi_N)] \mid_{X_N} = \mathcal{P}_N \mathcal{S} \mathcal{L} \mid_{X_N}$. Using the regularity of $\mathcal{L}$, we can find the following equivalent equation to (15) (see [4]):

\[ \phi = \mathcal{N} \phi + (I - \mathcal{P}_N) \mathcal{K} \phi = T \phi, \tag{16} \]

where $\mathcal{N} = \mathcal{P}_N - \mathcal{L}^{-1} \mathcal{P}_N (I - K)$ is a Newton–like iteration operator for (14) and the last part becomes small if $X_N$ is sufficiently large. The operator $\mathcal{N}$ is compact since it maps $X$ into the finite dimensional space $X_N$ and hence $T$ is compact. Due to the equivalence of (14) and (16), we have a compatible verification condition of the form: $T U \subset U$. And we split the candidate set $U$ into two parts: $U = U_N \cup U$, where $U_N \subset X_N$, $U_\ast \subset X_\sharp$ and $X_\sharp$ is the orthogonal complement of $X_N$ in $X$. Then we obtain the verification conditions.
Theorem 4 Let $U_N$, $U$, and $U'$ be sets defined above. If the following inclusions hold,
\begin{align}
NU & \subset U_N, \quad (17a) \\
(I - P_N)KU & \subset U', \quad (17b)
\end{align}
then there exists a fixed point $\tilde{\phi}$ of $T$ in $U$.

4. Computable verification conditions

Definitions of $K$ and $L$ enable us the calculation of $N\tilde{\phi}$: $N\tilde{\phi} = L^{-1}(L \tilde{\phi} - (\sum_{i} \psi_i \phi_i) \tilde{\phi}) = L^{-1}P_N S F_0 (\phi, \tilde{\phi})$, where $F_0(\phi, \tilde{\phi})$ is defined as: for $\psi \in X$, $(F_0(\phi, \tilde{\phi}), \psi) = (F(\phi, \phi), \psi) - (\nabla \tilde{\phi}, \nabla \phi) - (F' (\phi) \bar{\psi}, \phi)$. Then (17a) requires $\psi \equiv N\tilde{\phi} \in U_N$, and applying $L$ on both sides gives $P_N S F_0 \psi = L \psi = P_N S F_0 (\phi, \tilde{\phi})$. Using the projection property $P_N$, we can derive the following variational form: $(\nabla S L \psi, \nabla \phi) = (\nabla S F_0 (\phi, \tilde{\phi}), \nabla \phi)$ for $\psi \in X$. This can be written as $(L \psi, \phi) = (F(\phi, \phi), \phi)$ for $\psi \in X_N$. Note that $(L \phi, \phi) = (S^{-1} \phi - F(\phi, \phi) \phi) = (\nabla \phi, \nabla \phi) - (F' (\phi) \phi, \phi)$.

For given real numbers $c_0^i$, $\bar{c}_i^i$, $M_0$, and $M_1$, define $U_N = \{ \phi \in X_N | c_i^i \in [\bar{c}_i^i, c_i^i], i = 1, 2, 3, \alpha \in S N \cap T \}$ and
$U = \{ \phi \in X_N | \| \nabla \phi \|_0 \leq c_0 M_0, \| \nabla \phi \|_{\infty} \leq C M_0, \| \phi \|_{\infty} \leq C M_1, \| \phi \|_{\infty} \leq C M_1, \| \phi \|_{\infty} \leq C M_0 \}$ and set $U = U_N \cup U$.

For any $\phi \in X_N$, we can express it in terms of $\psi^i_N$: $\phi = \sum_{i=1}^{\infty} \psi^i_N$. Therefore, the verified solution of the above linear matrix problem assures the regularity of the matrix $J$ and the existence of the inverse of $L$. Set $\{ f_i \}$, $\{ \tilde{f}_i \}$, we then have the following Stokes equations:
\begin{align}
-\Delta \tilde{u} + \nabla \tilde{p} & = f(\phi, \tilde{\phi}) + \Delta u_N - \nabla p_N, \quad (19a) \\
\nabla \tilde{u} & = 0, \quad (19b) \\
-\Delta \tilde{\phi} & = g(\phi, \tilde{\phi}) + \Delta \phi_N. \quad (19c)
\end{align}

Using (8), (10) and (12) with (19), we get the following:
\begin{align}
\| (I - P_N) \tilde{u} \|_0 & \leq C_0^0 \| f(\phi, \tilde{\phi}) + \Delta u_N - \nabla p_N \|_0, \quad (20a) \\
\| \nabla \tilde{u} \|_0 & \leq C_0^0 \| f(\phi, \tilde{\phi}) + \Delta u_N - \nabla p_N \|_0, \quad (20b) \\
\| (I - Q_N) \tilde{\phi} \|_0 & \leq C_0^0 \| g(\phi, \tilde{\phi}) + \Delta \phi_N \|_0, \quad (20c) \\
\| \nabla (I - Q_N) \tilde{\phi} \|_0 & \leq C_0^0 \| g(\phi, \tilde{\phi}) + \Delta \phi_N \|_0, \quad (20d) \\
\| (I - P_N) \tilde{\phi} \|_0 & \leq C_3 \| f(\phi, \tilde{\phi}) + \Delta u_N - \nabla p_N \|_0. \quad (20e)
\end{align}

Hence, the verification condition (17b) holds if for $\tilde{\phi} \in U$
\begin{align}
M_f & = \| f(\phi, \tilde{\phi}) + \Delta u_N - \nabla p_N \|_0 \leq M_0, \quad (21a) \\
M_g & = \| g(\phi, \tilde{\phi}) + \Delta \phi_N \|_0 \leq M_1. \quad (21b)
\end{align}

5. Numerical results

In the verification step, interval arithmetic is used in the form of INTLAB, the interval arithmetic package for MATLAB on PC. The verification algorithm is basically the same as that in [4].

We select $a = 1/2 \sqrt{\Sigma}, b = \sqrt{\Sigma}$ and $P = 10$. Then the critical Rayleigh number $R_0 = 6.75$ can be attained at some special mode (see [3]).

Due to limitation of memories and slow performance, we verified each type near close to the critical Rayleigh number, e.g. $R/R_0 = 1$ etc. On the other hand, for roll type solutions, since some reduction of variables is possible, we succeeded to verify far from $R_0$, e.g., for $R/R_0 = 1.5$. Also we computed various kinds of approximate solutions developed sub–branches of each type such as 1 peak and 4 peaks branches, which showed shift of the critical Rayleigh numbers.

In the figures, the isothermal lines are drawn after adding the conduction solution, and note that stream lines for each type never change their shape during the change of relative Rayleigh numbers.

References


Figure 1: Isothermal lines of roll type solution for $R/R_c = 1.5$ on the plane $x_2 = 0$. Contour lines of speed and streamlines of it on the plane $x_2 = 0$.

Figure 2: Isothermal lines of rectangular type solution for $R/R_c = 1.5$ on the plane $x_3 = \pi/2$. Contour lines of speed and streamlines of it on the plane $x_3 = \pi$.

Figure 3: Isothermal lines of hexagonal type solution for $R/R_c = 1.5$ on the plane $x_3 = \pi/2$. Contour lines of speed and streamlines of it on the plane $x_3 = \pi$.

Figure 4: Isothermal lines of mixed type (rectangular and hexagonal) solution for $R/R_c = 1.5$ on the plane $x_3 = \pi/2$. Contour lines of speed and streamlines of it on the plane $x_3 = \pi$. 
A computer-assisted proof for the Kolmogorov flows of incompressible viscous fluid

Yoshitaka Watanabe†
†Computing and Communications Center, Kyushu University,
Fukuoka 812-8518, Japan
Email: watanabe@cc.kyushu-u.ac.jp

Abstract—A computer-assisted proof of non-trivial steady-state solutions for the Kolmogorov flows is described. The governing equation is the Navier-Stokes equation in a two-dimensional flat torus under a special driving force proposed by Kolmogorov. The verification method is based on the infinite-dimensional fixed-point theorem using Newton-like operator. This paper proposes a numerical verification algorithm which generates automatically on a computer a set including the exact non-trivial solution with local uniqueness. A discussed numerical example is taken into account of the effects of rounding errors in the floating point computations.

1. Introduction

Consider the following Navier-Stokes equations:

\begin{align}
\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} &= v \Delta u - \frac{1}{\rho} \frac{\partial p}{\partial x} + \gamma \sin \left(\frac{\pi y}{b}\right), \\
\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} &= v \Delta v - \frac{1}{\rho} \frac{\partial p}{\partial y}, \\
\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} &= 0,
\end{align}

where \((u, v, p)\) are velocity vector, mass density, and pressure and \(\gamma\) is kinematic viscosity, respectively. \(\gamma\) is a constant representing the strength of the sinusoidal outer force. Also \(\phi \Delta \phi = \frac{\partial^2}{\partial x^2} \phi + \frac{\partial^2}{\partial y^2} \phi\). The flow region is a rectangle \([-a, a] \times [-b, b]\) and the periodic boundary conditions are imposed in both directions. The aspect ratio is denoted by \(\alpha := b/a\).

The above equations (1–3) describe the Navier-Stokes flows in a two-dimensional flat torus under a special driving force proposed by Kolmogorov [1, 5] and have a basic solution which is written as

\((u, v, p) = (k \sin(\pi y/b), 0, d)\),

where \(k := b^2 \gamma / (\pi^2 \nu)\) and \(d\) is any constant. It is known that non-trivial solutions bifurcate from the basic solution at a certain Reynolds number, which is defined below, if and only if \(0 < \alpha < 1\) [1]. Okamoto-Shoji [5] computed numerically bifurcation diagrams with the Reynolds number as a bifurcation parameter varying the aspect ratio as a splitting parameter. They also strongly suggested stability for all \(0 < \alpha < 1\). Nagatou [2] took a new approach to this stability problem by employing the theory of verified computation and showed that the stability is mathematically rigorously assured for the cases of \(\alpha = 0.4, 0.7\), and 0.8. However, theoretical approach to the non-trivial solutions of the equations (1–3) has not been showed up to now.

The aim of this paper is to propose a method to prove the existence of the steady-state solutions of the Navier-Stokes equations (1–3) for given Reynolds number and aspect ratio by a computer-assisted proof. This method is based on the infinite-dimensional fixed-point theorem using Newton-like operator as well as on the spectral approximation and the constructive error estimates.

2. Nondimensionalization and function spaces

The letter \(T_\alpha\) denotes the rectangular region \((-\pi/\alpha, \pi/\alpha) \times (-\pi, \pi)\) for a given aspect ratio \(0 < \alpha < 1\). Introducing the stream function \(\phi\) satisfying \(u = \phi_y\) and \(v = -\phi_x\), so that \(u_x + v_y = 0\), the equations (1–3) can be rewritten as

\[ (\Delta \phi)_t - v \Delta^2 \phi - J(\phi, \Delta \phi) = \frac{\nu \pi}{b} \cos \left(\frac{\pi y}{b}\right) \]

by cross-differentiating equations (1) and (2) and eliminating the pressure \(p\). Here \(J\) is a bilinear form defined by

\[ J(u, v) := u_x v_y - u_y v_x. \]

The equation (4) is nondimensionalized using change of variables

\[ (x', y') = \left(\frac{\pi x}{b}, \frac{\pi y}{b}\right), \quad t' = \frac{\nu b}{\nu \pi}, \quad \phi'(t', x', y') = \frac{\nu \pi^3}{\gamma b^3} \phi(t, x, y), \]

and the Reynolds number \(R := \frac{\gamma b^3}{\nu \pi^3}\). After dropping the primes, an equation

\[ (\Delta \phi)_t - \frac{1}{R} \Delta^2 \phi - J(\phi, \Delta \phi) = \frac{1}{R} \cos(y) \]

is obtained.

We shall find steady-state solutions, where \((\Delta \phi)_t\) is equated to 0 in equation (6) in the region \(T_\alpha\), namely consider the following nonlinear problem:

\[ \Delta^2 \phi = -R J(\phi, \Delta \phi) - \cos(y) \text{ in } T_\alpha. \]
Assume that the stream function $\phi$ is periodic in $x$ and $y$, and the symmetric condition $\phi(x, y) = \phi(-x, -y)$ [2], then the equation (7) has a trivial solution $\phi = -\cos(y)$ for any $R > 0$. The aim of this paper is to verify the existence of non-trivial solutions by a computer.

From the assumptions imposed above, the solutions of eq. (7) should be obtained in the following function space $X^k \subset H^k(\Omega)$ ($k \geq 0$) such that

$$X^k := X^k_0 \oplus X^k_1 \oplus X^k_2 \oplus \cdots ,$$

$$X^k_0 := \left\{ \sum_{n=1}^{\infty} a_n \cos(\pi y) \left\{ \begin{array}{ll} a_n \in \mathbb{R}, \sum_{n=1}^{\infty} n^2 a_n^2 < \infty \end{array} \right\} \right\},$$

$$X^k_m := \left\{ \sum_{n=1}^{\infty} a_n \cos(\pi y) \left\{ \begin{array}{ll} a_n \in \mathbb{R}, \sum_{n=1}^{\infty} ((am)^2 + n^2)a_n^2 < \infty \end{array} \right\}, m \geq 1,$$

especially

$$X := X^3.$$

For all $\psi \in X^k$ can be represented by

$$\psi = \sum_{(m,n) \in Q} A_{mn} \cos(\pi x + n\pi y), A_{mn} \in \mathbb{R},$$

where

$$Q := \left\{ (m,n) \in \mathbb{N} \times \mathbb{N} \left| \begin{array}{l} m = 0 \text{ and } 1 \leq n \leq \infty \text{ or } \text{ or } 1 \leq m \leq \infty \text{ and } -\infty \leq n \leq \infty \end{array} \right\} ,$$

and it is noted that

$$\langle \cos(\pi x + n\pi y), \cos(k\pi x + l\pi y) \rangle \in \begin{cases} \frac{2\pi^2}{\alpha} & \text{if } k = m \text{ and } l = n \\ 0 & \text{else} \end{cases}$$

holds for any $(m,n),(k,l) \in Q$, where $(\cdot, \cdot)_{L^2}$ means the usual $L^2$-inner product in $T_\alpha$.

3. Approximate subspace and norm estimates

Let $X_N$ be the finite-dimensional subspace of $X$, which depends on an integer parameter $N$, defined by

$$X_N := \left\{ \sum_{(m,n) \in Q_N} A_{mn} \cos(\pi x + n\pi y) \left\{ A_{mn} \in \mathbb{R} \right\} \right\},$$

where

$$Q_N := \left\{ (m,n) \in \mathbb{N} \times \mathbb{N} \left| \begin{array}{l} m = 0 \text{ and } 1 \leq n \leq N' \text{ or } \text{ or } 1 \leq m \leq N \text{ and } -N \leq n \leq N' \end{array} \right\} .$$

Also let $X$ denote the orthogonal complement of $X_N$ in $X$ such that $X = X_N \oplus X_\perp$, then for any $\psi_\perp \in X_\perp$, can be represented by

$$\psi_\perp = \sum_{(m,n) \in Q} A_{mn} \cos(\pi x + n\pi y), A_{mn} \in \mathbb{R},$$

where $Q_\perp := Q - Q_N$. Now, defining the norm of $X$ as $\|\cdot\|_X := \|\cdot\|_{H^1(\Omega)}$, the following norm estimates hold.

Lemma 3.1 For any $\psi \in X$ and $\psi_\perp \in X_\perp$, it can be checked that

$$\|\psi\|_{L^2(T_\alpha)} \leq \alpha^{-\frac{3}{2}} \|\psi\|_X, \|\psi\|_{L^2(T_\alpha)} \leq \alpha^{-\frac{3}{2}} \|\psi\|_X,$$

$$\|\psi_x\|_{L^\infty(T_\alpha)} \leq C_3 \|\psi\|_X, \|\psi_x\|_{L^\infty(T_\alpha)} \leq C_3 \|\psi\|_X,$$

where

$$C_1 = \frac{1}{\alpha^3(\alpha + 1)^2}, \quad C_2 = \frac{1}{\alpha^3(\alpha + 1)^2}, \quad C_3 = \max \left\{ 1, \frac{2\sqrt{3}}{9\pi^2} \right\},$$

$$C_4 = \max \left\{ \frac{1}{(\alpha + 1)^2}, \frac{2\sqrt{7}}{9\alpha^2(\alpha + 1)^2} \right\},$$

$$C_5 = \frac{1}{\alpha(\alpha + 1)}, \quad C_6 = \max \left\{ 1, \frac{1}{2\alpha} \right\},$$

$$C_7 = \max \left\{ \frac{1}{\alpha + 1}, \frac{1}{2\alpha(\alpha + 1)} \right\}.$$

In actual calculations, $L^\infty$-estimates proposed by Plum [6] are also needed.

Lemma 3.2 (Plum,1992) For $\psi \in X$, the following assertion holds true:

$$\|\psi\|_{L^\infty(T_\alpha)} \leq C_8 \|\psi\|_{L^2(T_\alpha)} + C_9 \|\nabla \psi\|_{L^2(T_\alpha)} + C_{10} \|\Delta \psi\|_{L^2(T_\alpha)} ,$$

where $\|\cdot\|_{L^\infty(T_\alpha)}$ is the sup-norm and

$$C_8 = \frac{\sqrt{\pi}}{2\pi}, \quad C_9 = \frac{1.1548}{\sqrt{3} \sqrt{\frac{\alpha^2 + 1}{\alpha}}},$$

$$C_{10} = \frac{0.44722}{3} \frac{\sqrt{9\pi^4 + 10\alpha^2 + 9}}{5\alpha^3} .$$

Lemma 3.1 and Lemma 3.2 imply $L^\infty$-estimates immediately.

Lemma 3.3 For $\psi \in X$ and $\psi_\perp \in X_\perp$, the following estimates hold:

$$\|\psi\|_{L^\infty(T_\alpha)} \leq C_1 \|\psi\|_X, \|\psi\|_{L^\infty(T_\alpha)} \leq C_2 \|\psi\|_X,$$

$$\|\psi_x\|_{L^\infty(T_\alpha)} \leq C_3 \|\psi_x\|_X, \|\psi_x\|_{L^\infty(T_\alpha)} \leq C_3 \|\psi_x\|_X,$$

$$\|\psi_x\|_{L^\infty(T_\alpha)} \leq C_4 \|\psi_x\|_X, \|\psi_x\|_{L^\infty(T_\alpha)} \leq C_4 \|\psi_x\|_X,$$

$$\|\psi\|_{L^\infty(T_\alpha)} \leq C_5 \|\psi\|_X, \|\psi\|_{L^\infty(T_\alpha)} \leq C_5 \|\psi\|_X, \|\psi\|_{L^\infty(T_\alpha)} \leq C_5 \|\psi\|_X.$$

432
where
\[
C_{11} = \alpha^{-3}C_8 + \alpha^{-2}C_9 + \alpha^{-1}C_{10},
C_{12} = C_1C_8 + C_2C_9 + C_5C_{10},
C_{13} = \alpha^{-2}C_8 + \alpha^{-1}C_9 + C_{10},
C_{14} = C_2C_8 + C_5C_9 + C_{10},
C_{15} = C_3C_8 + C_5C_9 + C_{10},
C_{16} = C_4C_8 + C_5C_9 + C_{10}.
\]
Moreover, some “inverse” order estimates are needed.

**Lemma 3.4** For \(\psi_N \in X_N\) the following holds:
\[
\|\phi_N\|_{x} \leq C_{17}\|\Delta\phi_N\|_{L^2(T_0)},
\|\psi_N\|_{L^2(T_0)} \leq C_{18}\|\Delta\phi_N\|_{L^2(T_0)},
\|\psi_N\|_{L^2(T_0)} \leq C_{19}\|\Delta\phi_N\|_{L^2(T_0)},
\]
where
\[
C_{17} = N\sqrt{1 + \alpha^2},
C_{18} = \alpha N^2 \sqrt{1 + \alpha^2},
C_{19} = N^2 \sqrt{1 + \alpha^2}.
\]

4. Fixed-point formulation and error estimates

The bilinear form \(J\) has the following properties.
\[
(J(u), w)_{L^2} = (J(w), u)_{L^2} = -(J(u, w), v)_{L^2},
\quad u, v, w \in X^2,
J(u, v) \in X^0, \quad u, v \in X^1.
\]

Denote an approximate solution of eq.(7) by \(\phi_N \in X_N\) which is obtained by an approximate method. Then setting
\[
\psi := \phi - \phi_N
\]
and
\[
f(\psi) := -RJ(\phi_N + \psi, \Delta\phi_N + \Delta\psi) - \cos(\psi) - \Delta^2\psi_N,
\]
the problem (7) is rewritten as the residual form to find \(\psi \in X\) satisfying
\[
\Delta^2\psi = f(\psi) \quad \text{in} \ T_0.
\]
(17)

Note that \(\psi\) is expected to be small if \(\phi_N\) is an accurate approximation. By virtue of the property for \(J\) (14), \(f\) is the bounded continuous map from \(X\) to \(X^0\).

Moreover, it is easily shown that for all \(g \in X^0\), the linear problem \(\Delta^2\xi = g\) has a unique solution \(\xi \in X^4\). When this mapping is denoted by \(\xi = Kg\), setting the embedding from \(X^4\) into \(X^2\) by \(I\) and \(\Delta^{-2} := IK\), the operator \(\Delta^{-2} : X^0 \to X\) is a compact map because of the compactness of the embedding \(H^2(T_0) \hookrightarrow H^3(T_0)\) and the boundedness of \(K\). Therefore, eq.(17) is rewritten by a fixed-point equation:
\[
\psi = F\psi
\]
for the compact operator \(F := \Delta^{-2}f\) on \(X\).

Now, the \(H^2\)-projection \(P_N : X \to X_N\) is defined by
\[
(\Delta\psi - P_N\psi, \Delta\phi_N)_{L^2} = 0, \quad \forall\psi_N \in X_N.
\]
(19)

Note that for \(\psi = \sum_{m,n \in \mathbb{Q}} A_{mn} \cos(m\alpha + n\beta) \in X\) the projection coincides with truncation: \(P_N\psi = \sum_{m,n \in \mathbb{Q}} A_{mn} \cos(m\alpha + n\beta) \in X_N\). From this fact, the following constructive a priori error estimate is derived.

**Lemma 4.1** For each \(g \in X^0\), let \(\xi \in X^4\) be the solution of \(\Delta^2\xi = g\) and \(P_N\xi \in X_N\) be finite-dimensional approximation defined by eq.(19), then the following estimate holds.
\[
\|\xi - P_N\xi\|_{X^0} \leq C_5\|\xi\|_{L^2(T_0)}.
\]
(24)

Now, we apply the Newton-like method for nonlinear elliptic problems proposed by Nakao [3, 4] to the fixed-point equation (18). Using the projection \(P_N\), the fixed-point problem \(\psi = F\psi\) can be uniquely decomposed as the finite-dimensional (projection) part \(X_N\) and infinite-dimensional (error) part \(X\), as follows:
\[
\begin{cases}
P_N\psi = P_NF\psi, \\
(I - P_N)\psi = (I - P_N)F\psi,
\end{cases}
\]
(20)

where \(I\) is the identity map on \(X\). Suppose that the restriction of the operator \(P_N(I - F'(0)) : X \to X_N\) to \(X_N\) has an inverse
\[
[I - P_NF'(0)]^{-1}_{|_{X_N}} : X_N \to X_N,
\]
(21)

where \(F'(0)\) denotes the Fréchet derivative of \(F\). Note that this assumption is equivalent to the invertibility of a matrix, which is able to be numerically checked in actual verified computations (for example see Rump [7]). Applying the Newton-like method to the first term of eq. (20), the operator \(N : X \to X_N\) is defined by
\[
N\psi := P_N\psi - [I - P_NF'(0)]^{-1}_{|_{X_N}}P_N(\psi - F\psi),
\]
and also the compact map \(T : X \to X\) is defined by
\[
T\psi := N\psi + (I - P_N)F\psi.
\]
(22)

Then under the invertibility assumption of the existence for \([I - P_NF'(0)]^{-1}_{|_{X_N}}\), two fixed-point problems:
\[
\psi = T\psi
\]
and eq.(18) are equivalent. If the approximate solution \(\phi_N\) is sufficiently good, the finite-dimensional part of \(T\) will possibly be a contraction. On the other hand, the magnitude of the infinite-dimensional part of \(T\) is expected to be small when the truncation numbers of \(X_N\) are taken to be sufficiently large, because of Lemma 4.1.

The question which we must consider next is to find a solution of eq.(22) in a set \(U_N\), referred to as a candidate set. Let the finite-dimensional part of the candidate set \(U_N\) and the infinite-dimensional part of the candidate set \(U\), be balls with radius \(\gamma > 0\) and \(\beta > 0\) such as
\[
U_N := \{\psi_N \in X_N \mid \|\psi_N\| \leq \gamma\},
U := \{\psi \in X \mid \|\psi\| \leq \beta\},
\]
(23)

433
The candidate set $U \subset X$ is defined by
\[
U := U_N + U_*.
\] (25)

Now, the following verification condition is held [9].

**Theorem 4.1** Let $N'$ be the Fréchet derivative of $N$. For $Y_1, Y_2, Z_1(U)$ and $Z_2(U) > 0$ satisfying
\[
\|N'\|_X \leq Y_1, \quad \sup_{\phi_1, \phi_2 \in U} \|N' (\phi_1) \phi_2\|_X \leq Z_1(U),
\] (26)
\[
\|(I - P_N) F_0\|_X \leq Y_2,
\] (27)
\[
\sup_{\phi_1, \phi_2 \in U} \|(I - P_N) F'(\phi_1) \phi_2\|_X \leq Z_2(U),
\] (28)
if it hold that
\[
Y_1 + Z_1(U) < \gamma, \quad Y_2 + Z_2(U) < \beta,
\]
then there exists a fixed-point of $F$ in
\[
\hat{U} := \hat{U}_N + \hat{U}_*,
\]
\[
\hat{U}_N := \{ \psi \in X | \|\psi\|_X \leq Y_1 + Z_1(U) \},
\]
\[
\hat{U}_* := \{ \psi_* \in X | \|\psi_*\|_X \leq Y_2 + Z_2(U) \}.
\]

Moreover, this fixed-point is unique within the set $U$.

This theorem is proved by Banach’s fixed-point theorem. Note that in the references [9] the finite-dimensional part is taken to be a set of linear combinations of base functions with interval coefficients. Theorem 4.1 implies $L^\infty$-error estimates immediately.

**Theorem 4.2** Under the same assumption of Theorem 4.1, there exists the solution $\phi \in X$ of eq.(7) in
\[
\|\phi - \phi_N\|_{L^\infty(T_\gamma)} \leq C_{11}(Y_1 + Z_1(U)) + C_{12}(Y_2 + Z_2(U)).
\] (30)
Moreover, this solution is unique within
\[
\|\phi - \phi_N\|_{L^\infty(T_\gamma)} \leq C_{11}\gamma + C_{12}\beta.
\] (31)

**5. A verification result**

The interval arithmetic in each verification step was implemented using Sun ONE Studio 7, Compiler Collection Fortran 95 on FUJITSU PRIMEPOWER850 (CPU: SPARC64-GP 1.3GHz, OS: Solaris8). The approximate solutions were obtained by Newton-Raphson method using usual floating point arithmetic by double precision. The Reynold number is $R = 4$ and aspect ratio is $\alpha = 0.7$.

The verification algorithm executed successfully under the following values:
\[
Y_1 + Z_1(U) = 0.2104483239393 \times 10^{-9},
\]
\[
Y_2 + Z_2(U) = 0.1195514641468 \times 10^{-9},
\]
and we can assure that there exists non-trivial solution $\phi$ around the approximate solution $\phi_N$ bounded
\[
\|\phi - \phi_N\|_\infty \leq 0.9357796334788247 \times 10^{-9}
\]
with local uniqueness. Moreover the solution is unique for the bound of
\[
\|\phi - \phi_N\|_\infty \leq 1.2898092914771054 \times 10^{-5}.
\]

**References**


**Acknowledgments**

This work was supported by a Grant-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan (No.18540127, No.15204007, No.16104001).
Interplay of optical feedback and polarization dynamics in VCSELs

K. Panajotov* and H. Thienpont
Department of Applied Physics and Photonics, (TW-TONA), Vrije Universiteit Brussel (VUB), B-1050 Brussels (Belgium)

M. Sciamaenna
SUPELEC Campus de Metz 2 Rue Edouard Belin F-57070 Metz, France
(Dated: July 4, 2006)

We present experimental and theoretical studies of nonlinear dynamics induced by optical feedback in Vertical Cavity Surface Emitting Lasers (VCSELs) focusing on the impact of polarization mode competition. We unveil the bifurcation mechanism leading to multiple channels of interchanging linearly polarized (LP) steady states, study the way of their destabilization and the role of the external cavity modes and the polarization of light in this process. We distinguish two situations depending on the dichroism present in the VCSEL. In the first one, as typical for semiconductor lasers, with increasing the optical feedback strength the VCSEL exhibits a Hopf bifurcation from steady state to a limit cycle followed by a torus bifurcation to quasiperiodic behavior and then can become chaotic. Bistability between different single polarization dynamics: steady-state, periodic, quasi-periodic and chaotic dynamics is possible, however, the suppressed polarization mode is never excited. In the second case, when the VCSEL is biased in the region of polarization bistability, increasing the feedback strength can not only cause polarization switching but can also lead to chaotic emission in the two LP modes. Polarization switching can happen from a limit cycle in the x-LP mode to a steady state in the y-LP mode which then undergoes consecutive Hopf and torus bifurcations and becomes chaotic. Bistability between chaotic two LP mode emission and periodic single LP mode emission is demonstrated. We also consider the case of polarization selective optical feedback and the influence of the spontaneous emission noise or external noise in the system. For the last situation we demonstrate slow antithase hopping between the two LP modes together with rapid anticorrelated oscillations in the LP intensities at the external-cavity frequency. These last dynamics can be strongly enhanced by the noise leading to the phenomenon of coherence resonance. Finally, we consider the transition from long to short external cavity and show that the typical low-frequency fluctuation chaotic dynamics is anticipated by a regime of regular dynamics - fast pulses at the delay time modulated by a slow time periodic envelope. We discuss the interplay of optical feedback and polarization dynamics in this regime, too.

I. INTRODUCTION

Optical feedback (OF) in semiconductor lasers has attracted a lot of interest since more than twenty years both from fundamental point of view and from application side. OF introduces a time delay in the system (corresponding to the round-trip time $\tau$ in the external cavity (EC)) and therefore it results in a wealth of nonlinear dynamical effects. Depending on the ratio of the time-delay $\tau$ to the relaxation time of the laser $T_{RO}$ (determined by the photon - carrier coupling) three typical regimes can be distinguished: long EC regime when $T_{RO} < \tau$, short EC regime $T_{RO} \geq \tau$, and extremely short EC regime $T_{RO} \gg \tau$. The case of long EC with Edge Emitting Lasers (EEL) is the most extensively investigated. Very rich dynamics have been reported and classified in five typical dynamical regimes [1]. With increasing feedback strength these are: linewidth narrowing or broadening, external-cavity mode hopping, stable laser output and chaotic dynamics. Different routes to optical chaos, namely quasiperiodic [2] and period doubling [3] have been observed. Nowadays, OF induced dynamics in Vertical-Cavity Surface-Emitting Lasers or VCSEL is attracting more and more interest. VCSELs exhibits a number of properties superior to EEL[4], like circular beam shape, low threshold current and voltage and high speed of modulation, matrix fabrication and on wafer testing. However, VCSELs have the drawback of less stabilized polarization [5, 6]. Even when operating in the fundamental transverse mode only, VCSELs support two linear orthogonally polarized modes with slightly different frequencies and almost equal gains and losses [6]. These polarization degeneracy in VCSELs has major consequences in the case of OF. VCSELs are as sensitive to OF as the EEL in spite of their much higher output mirror reflectivity. Actually, this much higher reflectivity is compensated for by the much smaller intracavity round-trip time in VCSELs, so that the rate of the feedback is approximately the same as in EEL. Therefore, the same dynamical regimes that have been reported for EEL were experimentally observed in VCSELs [7]. As we mentioned, the existence of the two nearly degenerated linearly polarized (LP) modes in VCSELs makes these dynamical regimes more intricate. Polarization switching (PS) induced by polarized OF is

*Also at Institute of Solid State Physics, 72 Tzarigradsko Chaussee Blvd., 1784 Sofia, Bulgaria
observed experimentally even when the solitary device is with stable polarization [8]. Polarization isotropic OF can also lead to multistability and switching [9], giving rise to channelled light versus current characteristics, i.e. to polarization switchings that are periodical with current. Recently, polarization mode hopping accompanied by high frequency oscillations at the delay time has been demonstrated experimentally and numerically for fixed DC current at a point of polarization transition [10, 11]. In this paper we will present new studies on the dynamical leading to multiple channels of interchanging LP steady states, the way of their destabilization and the role of the EC modes and the polarization of light in this process.

II. RATE EQUATIONS FOR VCSEL SUBJECT TO WEAK OPTICAL FEEDBACK

OF introduces a time-delay in the laser system which can make its dynamical behavior complicated and its mathematical description difficult. Well accepted model that takes into account the OF as one-round trip delay contribution to the cavity internal field is the Lang-Kobayashi model [12]. Here we implement the Lang-Kobayashi model for the case of two mode laser, namely VCSEL working in a single transverse mode regime. In our model, we also include the gain nonlinearities that can be of different origin (like spectral and spatial hole burning, carrier heating, etc.). The spontaneous emission noise and its contribution to the laser linewidth through the amplitude-phase factor $\alpha$, typical for semiconductor lasers, is accounted for by two Langevin complex noise sources $F_{x,y}$ in the two field equations. In such a way, the phenomenological rate equations for VCSEL subject to a weak OF, read:

$$\frac{dE_x}{dt} = \frac{1}{\tau_{px}} \left[ \Gamma_x G_x - \frac{1}{\tau_{py}} \right] E_x + \kappa_x E_x (t - \tau) \exp(-j\omega_{px}\tau) + F_x,$$

$$\frac{dE_y}{dt} = \frac{1}{\tau_{py}} \left[ \Gamma_y G_y - \frac{1}{\tau_{px}} \right] E_y + \kappa_y E_y (t - \tau) \exp(-j\omega_{py}\tau) + F_y,$$

$$\frac{dN}{dt} = \frac{J}{e D \pi s^2} - \frac{N}{\tau_c} - G_x |E_x|^2 - G_y |E_y|^2.$$

Here $E_x$ and $E_y$ are the slowly varying x- and y- LP components of the electric field with optical frequency at threshold $\omega_{0x,y}$, $\alpha_{x,y}$ are their linewidth enhancement factors, $\Gamma_{x,y}$ are the confinement factors and $\tau_{px,y}$ are the corresponding photon lifetimes. The EC is characterized by its round-trip delay $\tau$, given by $\tau = 2L/c$ with $L$ being the EC length and $c$ the speed of light; and by the rate of the external feedback $\kappa_{x,y} = \eta_{\text{coup}}(1 - R) r_{x,y}^2 / (\tau_{in} \sqrt{R})$, with $R$ being the reflectivity of the VCSEL front mirror, $r_{x,y}$ the polarization dependent amplitude reflectivity of the external mirror and $\eta_{\text{coup}}$ the coupling efficiency in the VCSEL aperture and $\tau_{in}$ the photon round-trip time in the VCSEL cavity taken the same for the two polarization modes. In the rate equation for the carrier density

$$G_x = g_{yx} g_x (N - N_{tr}) \left( 1 - \varepsilon_{xx} |E_x|^2 - \varepsilon_{xy} |E_y|^2 \right),$$

$$G_y = g_{yx} g_y (N - N_{tr}) \left( 1 - \varepsilon_{x} |E_x|^2 - \varepsilon_{y} |E_y|^2 \right),$$

where $g_{yx, y} = C / n_{0x,y}$ are the group velocities, $N_{tr}$ is the carrier densities at transparency which is assumed to be the same for the x- and y- polarized modes and $g_{x,y}$ are the differential gains for the x and y linearly polarized mode. The gain compression is taken into account through $\varepsilon_{xx}$ and $\varepsilon_{xy}$ - the self gain saturation coefficients, and $\varepsilon_{x}$ and $\varepsilon_{y}$ - the cross saturation coefficients. The noise due to current injection has been neglected in the rate equation for the carrier density. In our model, we take into account the dependence of the cavity mode frequency on the injection current. Indeed, it is well known that VCSELS experience a frequency red shift with the injection current as it heats the cavity as flows through the doped Distributed Bragg Reflectors. In such a way, we have:

$$\omega_{0x,y} = \frac{n_0 \varepsilon_0 c}{n_{0x,y} L_c} \left( 1 - \frac{1}{n} \frac{dn}{dJ} \right).$$

Here, the refractive indices of the LP (x and y) modes are taken to vary linearly with the injection current with rate $dn/dJ$ and $n_{0x,y}$ are their values at zero injection current. Although the optical frequencies of the two LP modes are red shifted with current, the frequency difference between them is independent on $J$ in eqn.(6) as is usually the case in the experiments [6]. However, the gain difference between the two LP modes might depend on the current due to change in the LP mode frequency positions on the gain (absorption) curves. Assuming small changes we take linear dependence of gain difference:

$$g_y - g_x = \Delta g \left( \frac{J}{J_{PS}} - 1 \right),$$

where $J_{PS}$ is the current of polarization switching due to net gain equalization [6, 10].

III. BIFURCATION SCENARIOS IN VCSELS SUBJECT TO OPTICAL FEEDBACK

In general, PS in VCSELS can happen either through a region of polarization mode hopping or revealing a region of hysteresis [6]. In our model such situation is modelled phenomenologically with the gain equalization term [eqn. (7)]. We consider in this section the following VCSEL
FIG. 1: Bifurcation diagram, representing the maxima and minima of the LP resolved output power with bifurcation parameter injection current. The parameters are specified in the text and $J_{PS} = 3mA$. (a) no optical feedback: the solid black (gray) line represents the output power $P_x (P_y)$ of the x (y) LP mode; (b)-(d) optical feedback with $r_{ext} = 0.005$: the solid black (gray) line represents the LP output powers for increasing (decreasing) injection current. (b) is for the region of single polarization mode operation and (c) and (d) are for the region of polarization switching, representing $P_x$ and $P_y$, respectively.

parameters: $\alpha = 3$, $N_{ir} = 4 \times 10^6 \mu m^{-3}$, $\tau_p = 1.3 ps$, $\tau_e = 1 ns$, $L_c = 1.97 \mu m$, $m = 16$, $V = 2 \mu m^2$, $dn/dJ = 1.2 \times 10^{-3} \mu m A^{-1}$, $\Gamma = 0.06$, $v_g = 8.85 \times 10^4 \mu m ns$, $g_e = 4 \times 10^{-9} \mu m^{-2}$, $\Delta g = 1 \times 10^{-7} \mu m^{-2} / mA$, $\epsilon_s = 5 \times 10^{-6} \mu m^{-3}$, $\epsilon_c = 2 \epsilon_s$, $L = 20.4 cm$. The refractive indices for the two LP modes are slightly different due to residual birefringence in the VCSEL cavity; we take $n_{0x} = 3.5$ and $n_{0y} = 3.5001$. In such way, x (y) subscript corresponds to short (long) wavelength mode. VCSEL therefore experiences type one PS (from shorter to longer wavelength with increasing injection current). In the following we neglect the spontaneous emission noise because we are interested only in the deterministic polarization switching and instabilities and we consider polarization bistable solitary VCSEL, i.e. we set $J_{PS} = 3mA$ in Equation (7) (see Fig.1(a)). The solid black (gray) line in fig.1 (a) represents the output power $P_x (P_y)$ of the short (long) wavelength LP mode of the solitary VCSEL, i.e VCSEL without OF. In Fig. 1(b)-(c) we show the calculated bifurcation diagrams, i.e. the maxima and minima of the LP output power, with bifurcation parameter the injection current and with weak OF ($r_{ext} = 0.005$). Now, the solid black (gray) line represents the LP output powers for increasing (decreasing) injection current. The injection current, via the current induced heating and the corresponding refractive index and wavelength change, modifies the phase of the light fed back into the VCSEL - see Eqn.(6). Fig.1 (b) is for VCSEL biased in the region of single polarization mode operation and Fig.1 (c) and (d) represent $P_x$ and $P_y$, respectively, in the region of polarization switching. As can be seen from Fig.1 (b), when the gain difference between the two LP modes is large (the VCSEL is biased far from the polarization bistable region), the change of the VCSEL injection current periodically destabilizes the laser operation through a Hopf bifurcation to a limit cycle but do no cause polarization switching. The amplitude of the periodic oscillation in the limit cycle region first continuously increases and then decreases, forming a kind of a bubble. These bubbles appear almost periodically, separated by about $\Delta J \approx 0.007 mA$. The region of stable laser operation expands on the space of several limit cycle bubbles. As can be seen from Fig.1(b) the VCSEL remains in a single polarization state in the region of small and large (not shown) injection currents further from the polarization switching point. However, approaching the polarization switching point, as the gain difference between the two LP modes becomes small, the depressed polarization mode comes also into play. The reason for polarization switching to occur is the modulation with the current of the VCSEL mirror losses via the EC phase change induced by the injection current. This modulation is in general out of phase for the two LP VCSEL modes (with slightly different wavelengths) and can, therefore, compensate for their gain difference and make the VCSEL switch to the orthogonal LP mode - see Fig.1(c) and (d). The PS can take place from either a steady state or destabilized (oscillating) regime in one mode to a steady state in the orthogonal LP mode or

FIG. 2: Bifurcation diagram with bifurcation parameter the injection current for the case of polarization selective OF ($r_{ext} = 0.005$ and $r_{g} = 0.005$). (a) represent $P_x$ in the region of single polarization mode operation; (b) and (c) represent $P_x$ and $P_y$, respectively, in the region of polarization switching. The solid black (gray) line represents the corresponding LP output power for increasing (decreasing) injection current.
FIG. 3: Bifurcation diagram with bifurcation parameter injection current for the case of polarization selective OF ($r^x_{ext}=0.005$ and $r^y_{ext}=0$). (a) represent $P_x$ in the region of single polarization mode operation ($J=2.64mA$); (b) and (c) represent $P_x$ and $P_y$, respectively, in the region of polarization switching ($J=3.0mA$). The solid black (gray) line represents the corresponding LP output power for increasing (decreasing) injection current.

vice versa. In such a way, periodic regions of alternating lasing in the two LP modes are formed. As can also been seen from Figs. 1 (c) and (d) each of the multiple polarization switches defines a distinct region of bistability. The bistability is between steady states or steady state and periodic state depending on the injection current. The widths of these hysteresis regions is very small, around a $\mu A$, and in the experiments the bistability is masked by a random polarization mode hopping [10, 11].

It has been demonstrated experimentally that polarized OF can also cause polarization switching even when the solitary device is operating in a stable polarization [8]. We model the situation of polarized OF by using Equations (2)-(3) with either $k_x = 0$ or $k_y = 0$.

In Figs. 2 and 3 we show the calculated bifurcation diagrams as a function of the injection current for the situations of Fig.1 and polarization selective OF, i.e. $k_x > 0, k_y = 0$ and $k_x = 0, k_y > 0$, respectively. While for feedback in the lasing mode we observe similar to the case of polarization isotropic feedback sequence of periodic destabilization to limit cycle operation (Fig.2(a)) when the feedback is in the non-lasing mode no such destabilization is observed and the laser emits in a steady state (Fig.3(a)). When the VCSEL operates in the polarization bistable region, the periodic PS and channelled LI curve are again observed in the two cases of LP feedback. However, the LP mode which is submitted to OF is the only destabilized mode (compare Fig.2(b)(c) to Fig.3(b)(c)). Bistability between limit cycle in one LP mode and steady state in the orthogonal LP mode, or between two steady states is clearly visible in Figs. 2 and 3 ((b) and (c)).

IV. CONCLUSIONS

In summary, we have presented a numerical studies of bifurcation scenarios of optical feedback induced polarization switching and instabilities in VCSELs. Depending on the dichroism the VCSEL can either undergo a Hopf bifurcation followed by a torus bifurcation to quasiperiodic behavior and then chaotic behavior in a single linearly polarized mode or it can demonstrate polarization switching and chaotic emission in the two LP modes. Bistability between chaotic two LP mode emission and periodic single LP mode emission is also demonstrated. We also consider the case of polarization selective optical feedback and short external cavity; more results will be presented at the conference.

Acknowledgments

The authors acknowledge the support from the Inter-University Attraction Pole (IAP V/19) program of the Belgian government and the Fund for Scientific Research (FWO) - Flanders, Belgium. This work also benefited from project support from the Geconcertande Onderzoeks Aktie (GOA) and the Onderzoeksaanfd (OZR) of the Vrije Universiteit Brussels.

REFERENCES

Bifurcation to Polarization Chaos in VCSELs with Orthogonal Optical Injection

M. Sciamanna†, I. Gatare†,‡, H. Thienpont‡ and K. Panayotov†

† Supélec – Ecole Supérieure d’Electricité, Campus de Metz, and Laboratoire Matériaux Optiques, Photonique et Systèmes (LMOPS), common laboratory between the University of Metz and Supélec, CNRS UMR-7132
2 Rue Edouard Belin, F-57070 Metz, France
‡ Department of Applied Physics and Photonics (TW-TONA), Vrije Universiteit Brussel (VUB),
2 Pleinlaan, B-1050 Brussels (Belgium)
Email: marc.sciamanna@supelec.fr, gatarega_ign@metz.supelec.fr, hthienpo@tona.vub.ac.be
kpanajotov@tona.vub.ac.be

Abstract— In this paper we investigate nonlinear dynamics accompanying polarization switching (PS) in a vertical-cavity surface-emitting laser (VCSEL) submitted to orthogonal optical injection, i.e. the injected light is linearly polarized and orthogonal to that emitted by the free-running VCSEL. Qualitatively different dynamics are analyzed in the frequency detuning-injection strength plane. We experimentally report on rich dynamics such as injection locking, wave mixing, limit cycle, subharmonic resonance, and period doubling route to chaos. We complement our experimental results by a theoretical study. We find, in qualitative agreement with our experimental results, a period doubling cascade leading to a chaotic regime that involve the two orthogonal LP fundamental modes of the VCSEL. Finally, our study unveils the case of an elliptically polarized injection locking state that motivates further experimental investigations.

1. Introduction

Vertical-cavity surface-emitting lasers (VCSELs) have appeared as key components in several photonic applications, owing to their numerous advantages with respect to the conventional edge-emitting semiconductor lasers: low threshold current, circular beam with low divergence, high frequency modulation, on-wafer test and array capabilities [1]. However VCSELs may suffer from light polarization instabilities. The light emitted by a VCSEL is typically linearly polarized (LP) along one of two, orthogonal preferential directions but the light polarization may switch between these two LP modes as the injection current or temperature is varied [2], or when injecting external light into the VCSEL cavity [3]. It is crucial to understand the mechanisms determining the polarization stability of VCSELs, both in order to develop techniques for polarization control in polarization sensitive applications or, on the other hand, to make use of polarization instabilities for switching and routing applications.

Here we investigate a laser configuration in which a VCSEL is subject to optical injection from an external laser, and for which the injected light is linearly polarized and orthogonal to the LP light emitted by the free-running VCSEL. We also call this configuration “orthogonal optical injection”. As we vary the frequency detuning between master and slave lasers, or as we vary the injected power, we report experimentally on a rich polarization dynamical behavior. The slave VCSEL may exhibit polarization switching to the master polarization, which may be accompanied or not by injection locking. Nonlinear dynamics such as wave-mixing, limit cycle dynamics from Hopf bifurcation, period doubling and chaos, and subharmonic resonance are found in the plane of the injection parameters. Theoretical results from a rate equation model are in good qualitative agreement with experimental observations but also indicate new features not yet unveiled experimentally, such as a two-mode injection locking regime and the associated bifurcation scenario that leads to polarization switching to a single-mode injection locking state.

2. Experiment

2.1. Experimental setup

Our experimental investigations of polarization switching and dynamics induced by varying the intensity of the injected power for fixed detunings are conducted using the experimental setup shown in figure 1. We use a quantum-well VCSEL emitting around 845 nm as slave laser (SL). Its temperature and bias current are controlled by a laser driver CTR1. The injected light is provided by a tunable external cavity diode which is used as a master laser (ML). The tuning range of the ML wavelength is 845 to 855 nm and is controlled by the laser drive CTR2 by acting on the temperature or bias current of ML or even through piezoelectric control. The light emitted by the SL is collimated by a lens (COL) while the injected beam from the ML is focused on SL using another lens (L). The strength of the injected beam is changed using a polarizer P1. A half-wave plate, HWP1 and a second polarizer P2 are used to fix the linearly polarized injected beam to be orthogonal to the polarization direction of the VCSEL. An isolator ISO1 with 36-40 dB of attenuation plays the role of preventing ML from instabilities that may be induced by its own feedback effect or from the SL beam and other reflections from the detection branch. A 50/50 beam-splitter and a mirror M are used to align the SL and the ML with the detection branch. A half-wave plate HWP2 allows the choice of the polarization direction in which the measurements are carried out. The second isolator ISO2 with 36-40 dB of attenuation prevents the VCSEL from the light reflected by the fiber-coupling device FC. Spectral measurements are conducted using an optical spectrum analyzer or Fabry-Pérot spectrometer associated with a photodetector D1 and amplifier (Ampl) coupled to a computer (PC). The power emitted by SL or ML is measured using a powermeter PM.

2.2. Nonlinear dynamics accompanying polarization switching.

In order to investigate the dynamics induced by orthogonal optical injection, we bias the VCSEL with a drive current of 2.105 mA and the device temperature is fixed at 20°C. For this current, the free-running VCSEL emits in horizontal polarization and below the region of bistable polarization switching of the LI curve characteristic [4]. In figure 2, we show an experimental result of complicated dynamics that accompanies polarization switching induced by optical injection. The frequency detuning between the ML and the SL (\(\Delta \nu = \nu_{ML} - \nu_{SL}\)) is fixed at 2 GHz which is close to the frequency splitting between the two orthogonal polarized modes of the free-running VCSEL. Figure 2(a) corresponds to the injection locking regime, i.e., the VCSEL is locked to the ML frequency after a switch to the vertical polarization has been achieved. As the injection power is increased, the injection locking is exited through a Hopf bifurcation that is detected by the appearance of a limit cycle dynamic [see the two peaks on both sides of the main peak at the ML frequency in FIG. 2(b)]. The two small peaks in FIG 2(b) are separated from the main peak by a frequency splitting close to the frequency of the undamping relaxation oscillations (RO) of the free-running VCSEL (The measured value of the RO of the free-running VCSEL is 3.8 GHz). If the injection power is increased further, harmonics of the limit cycle are excited as shown in figure 2(c). By still increasing the injection power, the VCSEL undergoes a period doubling dynamics [FIG. 2(d)] leading to chaos which is characterized by the broadening of the VCSEL spectra with a large pedestal [FIG. 2(e)]. The presence of this chaotic attractor involves both orthogonal polarization modes, even if the vertically polarized mode is the dominating one. As it is shown in figure 2(f) and 2(g), for higher injection strengths, chaos is exited by a reverse period doubling bifurcation mechanism. For still strong injection levels, a second period doubling is again detected [FIG. 2(h)] which in turn evolves to a period one dynamics as we reach the maximum of the injection power.

FIG. 2. Polarization-resolved optical spectra for \(\Delta \nu = 2\) GHz and increasing the injection power \(P_{inj}\): (a) \(P_{inj} = 13.2\) µW; (b) 23.6 µW; (c) 36.6 µW; (d) 47.8 µW; (e) 80.3 µW; (f) 136.6 µW; (g) 204 µW; (h) 351 µW. The black (grey) line represents the vertical (horizontal) polarization. The vertical line shows the ML frequency.

FIG. 3. Polarization-resolved optical spectra of the VCSEL subject to optical injection at a frequency detuning of 10 GHz: (a) \(P_{inj} = 29\) µW; (b) 70 µW; (c) 251 µW; (d) 800 µW. (a), (b) and (c) shows subharmonic resonances with wave mixing and (d) represent a limit cycle dynamics (period-one). Only vertical polarization is represented. The vertical line shows the ML frequency.
In FIG. 3, we describe a scenario of the VCSEL dynamics for a detuning of 10 GHz. The injection power corresponding to FIG. 3 (a) is such that polarization switching (PS) is already achieved. However, unlike the case analyzed in FIG. 2, PS is not accompanied by injection locking to ML. The VCSEL emits in vertical polarization with a strong peak at the VCSEL frequency. Two side lobes are located on both sides of the VCSEL peak with a frequency splitting of 4 GHz which is close to the RO frequency. The strong excitation of the two side lobes that correspond to the RO undamping is found when the frequency splitting between ML and SL is close to the double of RO frequency. This type of dynamics is called subharmonic resonance and was also reported in edge-emitting lasers [5]. The other peak, which is located in the lower frequency range, is a result of wave mixing dynamics. As the injection strength is increased, other subharmonic resonances are also resolved [see FIG. 3(b) and 3(c)]. However, it is worth noting that an increase in the injection power leads to an increase in the frequency offset between the RO undamping and the VCSEL frequency. For higher injection levels, the subharmonic resonance disappears and the VCSEL dynamics is dominated by a limit cycle, period one dynamics [Figure 3 (d)].

3. Theory

3.1 The model

The model we use is built from the spin flip model (SFM) for the solitary VCSEL that operates in the fundamental transverse mode regime and which has been developed by San Miguel et al. [6]. In this model, the spin sublevels of the conduction and valence bands of a quantum-well VCSEL have been considered in order to account for the polarization properties of the laser electric field. We extend the SFM model to account for the injection of the external field. The rate equations for a VCSEL operating in the fundamental mode and submitted to orthogonal optical injection write:

\[
\frac{dE_n}{dt} = \kappa (1 + ia)(NE_n + inE_n, - E_n) - i(\gamma_s + \Delta \omega)E_n - \gamma \omega E_n \quad (1)
\]

\[
\frac{dE_y}{dt} = \kappa (1 - ia)(NE_n - inE_n, - E_n) + i(\gamma_s - \Delta \omega)E_n + \gamma \omega E_n + \kappa \omega E_n \quad (2)
\]

\[
\frac{dN}{dt} = \gamma N[1 + |E|^2 - |E_s|^2] - \mu + \eta \omega (E_x^* - E_y^*) \quad (3)
\]

\[
\frac{dn}{dt} = \gamma_n n[N[1 + |E|^2] + |E_s|^2] + \eta \omega (E_x^* - E_y^*) \quad (4)
\]

Where \( E_n, E_y \) represent the slowly varying amplitudes of the orthogonal linearly polarized (LP) components of the electric fields. The model considers two population inversions. \( N \) is the total population inversion between the conduction and the valence bands while \( n \) accounts for the population difference between carrier densities with opposite spins. Several parameters are considered: \( \kappa \) represent the decay rate of the electric field in the cavity, \( \gamma \) is the decay rate of the total carrier population inversion, and \( \gamma_s \) is a phenomenological modeling parameter that accounts for microscopic processes which lead to the mixing of carriers with opposite spin values. \( \alpha \) is the linewidth enhancement factor also present in the modeling of edge-emitting lasers and \( \mu \) is the normalized injection current (\( \mu \) takes the value 1 at threshold). The effect of linear amplitude anisotropy that considers the fact that the two LP modes may have different gain-to-loss ratio is modeled through \( \gamma_s \), while \( \gamma_p \) represents linear phase anisotropy or birefringence of the medium leading to frequency splitting between the two LP modes. The external optical injection is considered through \( \kappa \) which is the optical injection rate, \( E_{inj} \) which is the amplitude of the injected field and the frequency detuning \( \Delta \omega \) defined as the difference between the master laser frequency \( \omega_{inj} \) and an intermediate frequency \( \omega_{int} \) which is an average between the frequencies of the two orthogonal LP modes corresponding to x-and y-polarization directions, i.e. \( \omega_{inj} = (\omega_x + \omega_y)/2 \). Our numerical simulations have been conducted using the following choice for the VCSEL’s parameters: \( \gamma_s = 50 \text{ ns}^{-1}, \kappa = \kappa_{\omega} = 300 \text{ ns}^{-1}, \gamma_p = 30 \text{ rad.ns}^{-1}, \gamma_s = 0.5 \text{ ns}^{-1}, \alpha = 3, \mu = 1.5 \). These conditions lead to a free-running VCSEL emitting with a LP mode in the horizontal direction (x direction).

3.2 Bifurcation diagram and polarization dynamics

![FIG. 4. Mapping of different bifurcations showing qualitative change in the VCSEL dynamics in the injection parameter plane, i.e., frequency detuning vs. amplitude of the injected electric filed, \( E_{inj} \).](image)

In this section we complement our experimental study by a bifurcation analysis in order to better understand the mechanisms that underlie the nonlinear dynamics in a VCSEL subject to orthogonal optical injection. We use a continuation method to detect and follow the bifurcation points that characterize qualitative changes in the system dynamics for both stable and unstable solutions. In FIG 4 (a) we present the mapping of bifurcation boundaries in
the plane frequency detuning-injection field. We have resolved bifurcation that involves only the y-LP mode: one saddle-Node (SN\(y\)) and two Hopf bifurcation curves (\(H^x_y\) and \(H^y_y\)). SN\(y\) and \(H^y_y\) delimit a zone of injection locking of the y-LP mode to the ML frequency. In addition to the bifurcation of the steady state solutions, we have detected and have continued first-(1PD\(_y\)), second-(2PD\(_x\)) period doubling (PD), and even a torus (2T\(_x\)) bifurcation. Furthermore, in the range of negative detuning, we unveil another set of bifurcation lines that involve both x-LP and y-LP: A Saddle-Node bifurcation (SN\(_{xy}\)) and two Hopf bifurcation boundaries (\(H^x_{xy}\) and \(H^y_{xy}\)). SN\(_{xy}\) and \(H^y_{xy}\) bounds a thin region of Elliptically Polarized Injection-Locking (EPIL) regime for which both x-LP and y-LP are locked to the ML frequency.

In FIG. 5 we present polarization-resolved bifurcation diagrams showing the evolution of the maxima and minima of the intensities \(I_x\) and \(I_y\) of x-LP and y-LP modes respectively. The injection strength is increased from the free-running state (\(E_{inj} = 0\)) and for a fixed detuning. FIG. 5(a) represents a bifurcation scenario in the range of positive detuning (\(\Delta \omega = 42 \text{ rad/ns}\)). This allows exploring the zone delimited by the first period doubling bifurcation curve (1PD\(_x\)). As the injection strength is increased, a stationary x-LP solution is found until it bifurcates to a wave mixing regime at \(E_{inj} = 0.09\). By still increasing \(E_{inj}\), an abrupt polarization switching (PS) is achieved at \(E_{inj} = 0.013\) and the x-LP mode is then switched off. A further increased in \(E_{inj}\) shows that PS is followed by a period doubling cascade leading to chaos. It is worth noting that the resolved chaotic dynamics ranging from \(E_{inj} = 0.016\) to \(E_{inj} = 0.02\) involves both x- and y-LP modes. Therefore, this numerical result is in agreement with our experimental result presented in FIG. 2, where PS accompanied by period doubling route to chaos has been reported. For much stronger injection strengths, chaos evolves to periodic solutions of high period. A periodic window, which corresponds to a regime with low period, is also found around \(E_{inj} = 0.025\). FIG. 5(b) shows another interesting bifurcation scenario for a detuning of \(\Delta \omega = -30 \text{ rad/ns}\). As \(E_{inj}\) is increased, the laser undergoes a wave mixing dynamics that bifurcates, at \(E_{inj} = 0.03\), to the EPIL. The EPIL regime is exited through a Hopf bifurcation (\(E_{inj} = 0.032\)) that gives rise to a successive period doubling route to chaos. The chaos is then destabilized when PS is reached (\(E_{inj} = 0.052\)). For higher values of \(E_{inj}\) a steady y-LP injection-locked state in resolved and the x-LP is completely switched off.

4. Conclusions

In conclusion, we have experimentally investigated that PS in a VCSEL submitted to orthogonal optical injection may be accompanied by a period doubling route to chaos that involves both VCSEL’s LP fundamental modes. A subharmonic regime after PS has been also identified but only in the vertical LP mode. Our bifurcation analysis has shown a period doubling scenario (after PS) that qualitatively agrees with our experimental observations. Finally, we have reported on the EPIL regime and the associated bifurcations that motivate further experimental investigations.

5. Acknowledgments

This work has been supported by the IAP “Photon” Network of the Belgian government, the COST 288 European Action and the Region Lorraine (France).

References

Experimental observation of in-phase and anti-phase synchronization of chaos in coupled vertical-cavity surface-emitting lasers

Mitsutoshi Ozaki, Takaya Mihara, Hiroyuki Someya, Atsushi Uchida, and Shigeru Yoshimori

†Department of Electronics and Computer Systems, Takushoku University
815-1 Tatemachi, Hachioji, Tokyo 193-0985, Japan
Email: mitsutoshi@msn.com

Abstract – We experimentally observed chaos synchronization in mutually-coupled single-mode vertical-cavity surface-emitting lasers. In-phase and anti-phase synchronization is achieved for the polarization-resolved temporal waveforms between the two VCSELs.

1. Introduction

Vertical-cavity surface-emitting lasers (VCSELs) have been considered as novel devices for optical communications and wireless local area networks. Because of their short cavity length (a few micron) VCSELs emit in a single-longitudinal mode. Typical VCSELs have two-polarization modes and multi-transverse modes, which makes the dynamics very complicated [1,2]. However, basic dynamics of VCSELs with optical feedback are very similar to the edge emitting semiconductor lasers, such as low frequency fluctuations (LFF) and coherence collapse. VCSELs are sensitive to optical injection or optical feedback due to their short cavity length, in spite of high reflectivity of their facets (more than 99%).

In this study, we experimentally observe synchronization of chaos in two mutually-coupled VCSELs. Synchronization phenomena in VCSELs have already been reported [3-5]. However, the detail investigation is required for the purpose of communication applications using chaos in VCSELs. We observe the polarization-resolved temporal dynamics between the two VCSELs.

2. Vertical-cavity surface-emitting lasers (VCSEL)

The structure of VCSELs is shown in Fig. 1. The laser light emits to the perpendicular direction to the surface. VCSELs have some advantages compared with edge-emitting conventional semiconductor lasers: low threshold current, compact, high efficiency, large modulation bandwidth, and wafer-scale integration capability for large array configuration.

3. Experiment

3.1. Polarization dynamics in one VCSEL

Figure 2 shows our experimental setup for the observation of the dynamics of two polarization modes. An external mirror is set in front of a VCSEL to obtain self-feedback laser light. Chaos is induced by time-delayed optical feedback. The polarization modes of the VCSEL are resolved into two orthogonal components (x- and y-modes) by using a polarizer. The two polarization-mode dynamics are observed by an oscilloscope through an
photodetector. Figure 3 shows the temporal waveforms of the two polarization modes and Figure 4 shows the corresponding radio-frequency (RF) spectra of the two modes. The two chaotic temporal waveforms are anti-correlated, which is called anti-phase dynamics. The RF spectra show that the fundamental frequency of the chaotic waveforms is 170 MHz, which corresponds to the inverse of the round-trip time (5.88 ns) from the VCSEL to the external mirror.

3.2. Chaos synchronization in mutually coupled VCSELs

To observe chaos synchronization, the two VCSELs are mutually coupled to each other without the external mirror. Figure 5 shows the experimental setup for chaos synchronization. We set the injection currents of 2.50 and 2.09 mA for VCSEL 1 and 2, respectively. The coupling time (round-trip) between the two VCSELs is 6.41 ns. A half wave-plate is used to match the polarization direction between the two VCSELs. The temporal dynamics of the two VCSELs are observed by using a digital oscilloscope through two photodetectors. The optical wavelengths of the two VCSELs are controlled by changing the temperature of the VCSELs.

We use injection locking to achieve chaos synchronization. Injection locking is a technique to match the optical wavelength between two lasers by injecting a laser beam. Injection locking is achieved when the temperature of the VCSELs are precisely controlled in our experiment. The temperatures are set to 293.00 and 301.67 K for VCSEL 1 and 2, respectively. Using injection locking, the wavelength of the two VCSELs completely matches to 858.036nm as shown in Fig. 6.
Synchronization of chaos is observed between the x-modes of the two VCSELs as shown in Fig. 7. The temporal waveforms are in-phase oscillations. Moreover, the RF spectra of x-modes of two VCSELs are shown in Fig. 8. The RF spectra between two VCSELs matches well. The peak frequency appears at 156 MHz, corresponding to the inverse of the round-trip time (6.41 ns) between the two VCSELs. Figure 9 shows the correlation plots of the two temporal waveforms of x-mode in the two VCSELs. The cross correlation function is 0.943.

Synchronization of chaos is observed between the x-modes of the two VCSELs as shown in Fig. 7. The temporal waveforms are in-phase oscillations. Moreover, the RF spectra of x-modes of two VCSELs are shown in Fig. 8. The RF spectra between two VCSELs matches well. The peak frequency appears at 156 MHz, corresponding to the inverse of the round-trip time (6.41 ns) between the two VCSELs. Figure 9 shows the correlation plots of the two temporal waveforms of x-mode in the two VCSELs. The cross correlation function is 0.943.
We also observe chaos synchronization between different modes in the two VCSELs. Figure 10 shows the temporal waveforms of x-mode of VCSEL 1 and that of y-mode of VCSEL 2. The temporal waveforms are anti-correlated and anti-phase synchronization of chaos is observed. We found that anti-phase dynamics in VCSELs are responsible for anti-phase synchronization. Figure 11 shows the correlation plot between x-mode of VCSEL 1 and y-mode of VCSEL 2. The cross correlation function is -0.942. We thus observe both in-phase and anti-phase synchronization of chaos between the polarization-resolved temporal waveforms in mutually-coupled VCSELs.

3. Conclusion

We have observed synchronization of chaos in mutually-coupled single-mode VCSELs. The polarization-mode dynamics are observed in the presence of the external mirror. When the polarization modes are resolved into two orthogonal components (x-mode and y-mode) by using a polarizer, we found that chaotic anti-phase dynamics is observed between the two polarization modes.

In-phase synchronization of chaos is achieved between the same polarization modes of the two VCSELs. We also observe anti-phase synchronization between the different polarization modes because of anti-phase dynamics. We measure cross correlation to measure the accuracy of synchronization. Good cross correlation of 0.943 and -0.942 is obtained for the same and different modes, respectively.

References
Chaos Synchronization in Modulated Lasers

R. Meucci †, F. Salvadori ‡, K. Al Naimee †, ∗, and F. T. Arecchi ‡
†CNR - Istituto Nazionale di Ottica Applicata, Florence, Italy
‡Dept. of Physics, University of Florence, Italy
* Dept. of Physics, College of Sciences, University of Baghdad, Iraq

Abstract—We present experimental evidence of control and synchronization of crisis-induced intermittency regimes in modulated lasers. These bursting regimes are characterized by the alternation between a small amplitude chaotic regime and high amplitude periodic orbits originated by saddle node bifurcations. Control of bursting is achieved by means of a feedback method while synchronization is obtained by using a bidirectional coupling scheme.

1. Introduction

In a chaotic system it is relatively easy to observe an interior crisis [1, 2]. In such a case, a chaotic attractor suddenly expands in phase space including the region of an unstable orbit. Near after the crisis, whatever is the initial condition, the system visits the whole phase space of two dynamical objects, even though for different amounts of time. Any initial condition generates a trajectory visiting a single attractor; however, a memory of the previously separated regions remains and the fractional time spent on the formerly unstable orbit results in bursts of anomalous amplitude signal (crisis-induced intermittency [3]). In this work, we focus the attention on experiments on controlling bursting in a periodically driven CO₂ laser as well as on synchronization between two lasers exhibiting this regime.

2. Control of bursting

We consider a single-mode CO₂ laser with an intracavity electro-optic loss modulator (EOM)(see Fig.1). The cavity length is L = 1.42 m and the total transmission T is about 0.10 for a single pass. The intensity decay rate can be expressed as

\[ k(t) = k_0(1 + \alpha \sin^2(F_{\text{mod}}(t) + B_0)) \]  

(1)

where \( k = cT/L \), \( c \) being the speed of light in the vacuum, \( \alpha = (1 - 2T)/2T \), \( B_0 \) is a bias voltage and \( F_{\text{mod}}(t) \) is the modulation applied to the EOM. We consider sinusoidal modulation \( F_{\text{mod}}(t) = A \sin(2\pi f t) \) where \( f = 100 k\text{Hz} \) is about twice the relaxation frequency of the laser [4].

The modulation signal is provided by a waveform generator (WG). The laser is pumped by a DC discharge current stabilized at 8.00 mA, while the threshold current is 6.50 mA. The filter NF (Fig. 1(b)) has a notch frequency given by \( f_{\text{notch}} = 1/(2\pi \sqrt{L_2 C_2}) \), its input is the laser intensity and its output is amplified by \( a \). Increasing the modulation amplitude \( A \) (see Fig. 2(a)), the system undergoes a sequence of subharmonic bifurcations leading to a small amplitude chaotic attractor. A further increase of \( A \) (above 1.2 V) induces the occurrence of a regime where bursts of high-amplitude orbits of period three and period four (P3 and P4) are intercalated with the small amplitude chaotic attractor (Fig.2(b) and (c)).

At the interior crisis, the chaotic attractor expands into the region of the unstable P3, thus giving rise to a new attractor region which can be still discriminated by its frequency content. If we want to quench this region, we send its main frequency components (P3 and P4) as a negative feedback signal via the frequency filter. If on the contrary we change the feedback sign, we enhance the role of this region against the rest of the attractor. The frequency components corresponding to the subharmonic attractor are slightly affected by the filter (NF) whose notch frequency is selected at the first subharmonic \( f/2 \), that is, 50 kHz. The filter is AC coupled, so that the low frequency components, which control the long time dynamics, are not affected. The filter output is amplified and sent to the negative or positive
input of the differential amplifier (R). When the control is inserted, the modulation $F_{\text{mod}}(t)$ is perturbed by $aV_{\text{out}}$, where $V_{\text{out}}$ and $a$ are the output of the filter and its amplification respectively. We define the coupling strength $G$ as the ratio $aV_{\text{out}}/A$ between the perturbation and the modulation.

In Fig. 3 we report the experimental data at $A = 1.25\, V$, for the uncontrolled dynamics (a,d) and the for the controlled one, both with negative (b,e) and positive (c,f) feedback. On the left we plot the frequency spectra and on the right the corresponding attractors, reconstructed by the embedding technique. The larger amplitude chaotic attractor is drastically reduced in (b,e) giving evidence of burst suppression; on the contrary it is enhanced for a positive feedback in (c,f).

A quantitative evidence of suppression of the large amplitude bursts is provided by the fractional time indicator which is measured by:

$$\eta = \frac{\left(\sum T_{\text{BK}}\right)}{T}$$

(2)

$T$ being the duration of the recorded time series and $\sum T_{\text{BK}}$ the total duration of burst events within $T$. The bursts are detected by selecting a threshold in order to distinguish the large amplitude spikes from the small amplitude chaotic dynamics. The behaviour of $\eta$ from the non-controlled to the controlled regime is reported in Fig. 4 as a function of the coupling strength $G$. The transition toward complete control ($\eta = 0$) is characterized by a power law with an exponent $-1/2$

$$\eta \propto |G - G_{c}|^{-1/2}$$

(3)

Such a behaviour, confirmed also by numerical simulations, indicates the presence of a phase transition of the type I intermittency [5].
3. Synchronization of bursting

Chaos synchronization in lasers is one of the most active fields investigated in nonlinear dynamics. During the last decade many important aspects of this problem have been studied using CO$_2$ lasers [6, 7, 8]. In these experimental configurations, a saturable absorber is introduced inside the laser cavity inducing self-sustained chaotic pulsation regime known as passive Q-switching (PQS). A similar dynamic, obtained by using an intracavity electro-optic modulator, has been investigated in Ref.[9], where the effects of noise on synchronization of homoclinic chaos have been numerically and experimentally provided. Here, we focus on synchronization of bursting behavior in two non-identical lasers.

The experimental setup shown in Fig.5 consists of two single-mode lasers with intracavity electro-optic modulator. The cavity length is $L_1 = 1.43\, m$ and the total transmission is $T_1 = 0.10$ for the first laser; $L_2 = 1.35\, m$ and $T_2 = 0.090$ are the parameters for the second one. The decay rates are expressed as $k_{1,2} = (cT_{1,2})/(L_{1,2})$. Both lasers are driven by a sinusoidal signal $A \sin(2 \pi f t)$ provided by two phase-locked oscillators at $f = 100\, kHz$. The lasers are bidirectionally coupled through their intensity signal difference used as modulation amplitude of $A$

$$F_{1 mod}^1 = A_1(1 + \epsilon(y_1 - x_1)) \sin(2\pi ft)$$
$$F_{2 mod}^2 = A_2(1 + \epsilon(x_1 - y_1)) \sin(2\pi ft)$$

The parameters of both lasers are set just after the onset of the interior crisis, where they display intermittent behavior as shown in Fig.6.

As the coupling amplitude value $\epsilon$ is suitably adjusted, the jumps on the unstable orbits occur synchronously (Fig.7). A more detailed analysis reveals that the two lasers are nearly phase synchronized during the small amplitude chaotic regime while during the bursts their dynamics is anti correlated. Globally in the bursting behavior the lasers display frequency synchronization. Such a behavior is clearly shown when we report the inter-burst time intervals in $x - y$ representation.

In conclusion, we have demonstrated that the crisis-induced intermittency regime can be controlled by means of a feedback technique using a passive filter. Furthermore, this dynamical regime, largely studied in neuroscience, can be synchronized in two bidirectionally coupled lasers.

Acknowledgments

Work partly supported by FIRB contract No.RBAU01B49F-002.

References


Chaotic dynamics and synchronization in microchip lasers coupled with opto-electronic feedback

Keisuke Mizumura, Atsushi Uchida, and Shigeru Yoshimori

Department of Electronics and Computer Systems, Takushoku University
815-1 Tatemachi, Hachioji, Tokyo 193-0985, Japan
Email: a-uchida@es.takushoku-u.ac.jp

Abstract – We experimentally observe chaotic behavior of laser output in two Nd:YVO₄ microchip solid-state lasers coupled mutually with opto-electronic feedback. The output of one laser is detected by a photodiode and the electronic signal converted from the laser output is fed back to the pumping for the other laser. Chaotic fluctuation of the laser output is observed when the relaxation oscillation frequency is close to each other between the two lasers. Synchronization of periodic waveform is also obtained when the lasers have a single-longitudinal mode.

1. Introduction

Chaotic dynamics and synchronization behavior have been investigated in many laser systems for its potential applications to secure communications and spread spectrum communications. A variety of temporal dynamics with high dimensionality can be generated in lasers with time-delayed feedback. Semiconductor lasers with delayed optical feedback have been found to generate chaotic oscillations, but these systems are sensitive to optical phase variations when the feedback optical field is coherent with optical field inside the laser cavity. Compared with optical feedback, opto-electronic feedback is reliable because the system is insensitive to optical phase variations [1]. Chaos in semiconductor laser with opto-electronic feedback has been reported and quasiperiodic route to chaos has been observed [1]. Synchronization of chaos in one-way and mutually coupled semiconductor lasers with opto-electronic feedback has also been reported [2,3]. The use of cross-coupled semiconductor lasers with opto-electronic feedback has been pointed out for the prediction of the system dynamics of epidemics [4]. Chaotic dynamics and phase synchronization has also been reported in CO₂ gas lasers with opto-electronic feedback [5,6].

Although the opto-electronic feedback method has been applied to semiconductor lasers and gas lasers for generating chaos, there have been few reports on solid-state lasers with opto-electronic feedback. Solid-state lasers that have a short cavity length (typically less than one millimeter) are called microchip lasers. Due to their short cavity length, lasing occurs in a single mode or in a few longitudinal modes. The dynamics of microchip solid-state lasers have been investigated numerically and experimentally [7,8]. Most studies are focused on chaos induced by pump modulation and loss modulation in microchip lasers. A few reports on chaos in microchip lasers with opto-electronic feedback to the intracavity loss modulator has been reported [9-11]. The dynamics of microchip lasers with opto-electronic feedback may provide a different aspect from the case of semiconductor lasers and gas lasers.

In this study we investigate the chaotic dynamics of a Nd:YVO₄ microchip solid-state laser with opto-electronic feedback to pumping. The relaxation oscillation frequency of the microchip laser and the feedback gain are changed. We mutually couple two microchip lasers with opto-electronic feedback and observe the chaotic dynamics. Synchronization phenomenon is also observed.

2. One laser with opto-electronic feedback

2.1. Experimental setup

Figure 1 shows our experimental setup for the observation of chaotic dynamics in a microchip laser with opto-electronic feedback. We used a Nd:YVO₄ microchip crystal, pumped by a laser diode with two focusing lenses. The microchip laser oscillated with two-longitudinal modes at a wide range of pumping power.
The temperatures of the microchip crystal and the laser diode were controlled by thermo-electric coolers (resolution of 0.01 K) for fine tuning of the laser frequencies. The output of the microchip laser was detected by a photodiode and converted to an electronic signal. The electronic signal is fed back to the injection current of the pumping laser diode through a DC block (high-pass filter with a cut-off frequency of 16 kHz), an attenuator, and an electronic amplifier (bandwidth from DC to 10 MHz, Gain of 40 dB). Coaxial attenuators were combined to change the attenuation ratio from -40 to -20 dB and the net gain of the whole feedback loop was changed from -4 to 16 dB. The feedback delay-time was 0.25 μs, which was comparable with the inverse of the relaxation oscillation frequency of the microchip laser ($f_r = 1.4$ MHz, $1/f_r = 0.714$ μs). The positive feedback is achieved. Chaotic temporal waveforms of the microchip laser were detected with a photodiode and a digital oscilloscope. Radio-frequency (RF) spectrum of the laser output was measured with a RF spectrum analyzer.

2.2. Experimental results

We fixed the net gain of the feedback loop of 7 dB and varied the relaxation oscillation frequency by changing the injection current of the pumping laser diode. As the relaxation oscillation frequency $f_r$ (injection current $I$) is increased the temporal dynamics is changed. Figure 2 shows the temporal waveforms and the corresponding RF spectra at various relaxation oscillation frequencies. Switching behavior is observed just above the laser threshold ($I = 345$ mA) and sustained relaxation oscillation is observed at around $f_r = 2.00$ MHz ($I = 360$ mA). Period-1 oscillation is observed at $f_r = 2.22$ MHz ($I = 380$ mA) as shown in Figs. 2(a) and 2(b). As the relaxation oscillation frequency is increased, the coexisting state of period-1 and chaos is observed at $f_r = 2.27$ MHz ($I = 393$ mA) as shown in Figs. 2(c) and 2(d). Chaotic oscillation appears from $f_r = 2.30$ MHz to $f_r = 3.54$ MHz (from $I = 398$ mA to $I = 517$ mA). The transition from period-1 to chaos can be observed. We cannot find a clear bifurcation scenario such as period-doubling and quasiperiodic breakdown routes.

We changed both the relaxation oscillation frequency $f_r$ and the feedback gain $G$ simultaneously and create a two-dimensional dynamical map. Figure 3 shows the dynamical map as functions of the relaxation oscillation frequency and the feedback gain. Below $f_r = 2.0$ MHz, laser output shows relaxation oscillation at small $G$ and switching behavior at large $G$. At around $f_r = 2.5$ - 3.5 MHz, the transition from period-1 to chaos through coexisting state is observed. Chaos can be observed at wide range of the parameter space at large $G$. At $f_r > 3.5$ MHz chaotic oscillation disappears and relaxation oscillation is observed. The chaotic region of $f_r = 2.5$ - 3.5 MHz closely corresponds to the inverse of the feedback delay time ($f_r = 4.0$ MHz, $\tau = 0.25$ μs). Therefore the chaotic dynamics occurs because of the time-delayed feedback effect.
3. Two lasers coupled with opto-electronic feedback

3.1. Experimental setup

We investigate the dynamics of mutually coupled two microchip lasers with opto-electronic feedback. Figure 4 shows the experimental setup. The output of one Nd:YVO₄ microchip laser (Laser 1) is detected by a photodiode and the converted electronic signal is injected into the injection current of a pumping laser diode for the other Nd:YVO₄ microchip laser (Laser 2) through a DC block, an attenuator, and an amplifier (same components in the previous section). The output of Laser 2 is also detected with another photodiode and the electronic signal is injected into the injection current of the other laser diode for Laser 1. The cross coupling through opto-electronic loop provides complex dynamical behavior.

3.2. Experimental results

We fixed the feedback gains of $G_1 = 0$ and $G_2 = 8$ dB for Laser 1 and 2, respectively, and the relaxation oscillation frequency of $f_{r2} = 2.32$ MHz for Laser 2. We changed the relaxation oscillation frequency for Laser 1 to observe the dynamics of the mutually coupled microchip lasers. When the relaxation oscillation is less than 0.66 MHz ($I_1 < 320$ mA), slow switching behavior between no lasing state and a constant output is observed. The switching frequency is tens of Hz depending on the injection current. As the relaxation oscillation frequency is increased, period-1 oscillation is observed at the range of $0.66 < f_{r1} < 1.94$ MHz ($320 < I_1 < 367$ mA), as shown in Figs. 5(a) and 5(b). Sustained relaxation oscillation is also observed from 1.94 $< f_{r1} < 2.02$ MHz ($367 < I_1 < 374$ mA). Then quasiperiodic oscillation appears from 2.02 $< f_{r1} < 2.12$ MHz ($374 < I_1 < 380$ mA) because of the existence of the two relaxation oscillation frequencies of $f_{r1}$ and $f_{r2}$, as shown in Figs. 5(c) and 5(d). As $f_{r1}$ approaches $f_{r2} = 2.32$ MHz, chaotic oscillation is observed at the range from 2.12 $< f_{r1} < 2.68$ MHz ($380 < I_1 < 412$ mA), as shown in Figs. 5(e) and 5(f). The rf spectrum is broadened in the presence of chaos. Quasiperiodic oscillation is observed when $f_{r1}$ is further increased. Finally sustained relaxation oscillation appears at $f_{r1} > 2.78$ MHz ($I_1 > 416$ mA).

We next changed both $f_{r1}$ and $G_1$ simultaneously and create a two-dimensional dynamical map. We fixed $f_{r2} = 2.32$ MHz and $G_2 = 8$ dB, respectively. Figure 6 shows the two-dimensional map of the dynamics of Laser 1 and 2 as functions of $f_{r1}$ and $G_1$. Chaos is observed when $f_{r1}$ is close to $f_{r2} = 2.32$ MHz because of the nonlinear interaction between $f_{r1}$ and $f_{r2}$. Quasiperiodic breakdown route to chaos is observed. It is note that the dynamics of Laser 1 and 2 are slightly different, where period-1 oscillation only appears for Laser 1. This asymmetry behavior results from the mismatch of the two laser parameters.
4. Synchronization of periodic waveforms

We investigate synchronization in the mutually coupled microchip lasers with opto-electronic feedback. We decrease the injection current in order to achieve single-longitudinal-mode operation of the microchip lasers. The injection currents for the two lasers are set to $I_1 = 326$ mA and $I_2 = 320$ mA, respectively. The relaxation oscillation frequency is matched between Laser 1 and 2 ($f_{r1} = f_{r2} = 1.17$ MHz). We set the feedback gain of $G_1 = -2$ dB and $G_2 = 0$ dB. Figure 7 shows the temporal waveforms of the two microchip lasers. The period-1 oscillations are synchronized to each other with time delay. The time-shifted cross correlation shown in Fig. 7(b) indicates synchronization of the periodic waveforms.

5. Conclusion

We have experimentally observed the dynamics of Nd:YVO₄ microchip lasers with opto-electronic feedback. The laser output is detected and fed back to the injection current of the laser diode for pumping. Chaotic oscillations are observed in a microchip laser with opto-electronic self-feedback. Two microchip lasers are coupled with the opto-electronic signal obtained from the other microchip laser. Quasiperiodic breakdown route to chaos is observed. Synchronization of period-1 waveforms is also observed.

References
Recent developments in topological-computational approaches to nonlinear dynamics: an overview

Hiroshi Kokubu†

†Department of Mathematics, Kyoto University
Kyoto, 606-8502, Japan
Email: kokubu@math.kyoto-u.ac.jp

Abstract—This paper serves as an overview of the session, giving a brief summary of the recent developments of topological-computational approach in the study of nonlinear dynamics and bifurcations. It also includes some recent results of the author on the dynamics in a slow-fast system of ODEs.

1. Topological-computational approaches to nonlinear dynamics

In spite of progress in the theoretical study of nonlinear, especially chaotic, dynamical systems, one often encounter additional difficulties in verifying required conditions in order to apply a general theory to a concrete nonlinear system that models an interesting nonlinear phenomenon. In many cases, dynamical systems are given in the form of differential or difference equations, but analytical methods are usually very limited for such verification purposes. On the other hand, naive numerical simulation is not always trustworthy, because what one sees in a finite time numerical computation can be very different from the true asymptotic behavior of the system; it is now broadly known as a characteristic of chaotic systems that very small computation errors can lead to large differences in the long-term behavior of solutions.

Recent developments of numerical computation and visualization techniques, however, provide new methods, with an aid of computers, for studying nonlinear dynamical systems, in particular their global structures and bifurcations, such as high-quality visualization of invariant manifolds and sets or bifurcation diagrams in three (or even higher) dimensions, bounding errors in numerical integration of ODEs by using the interval arithmetic software that enables us to rigorously follow trajectories of dynamical systems, etc. In particular, the rigorous numerics combined with some topological methods are successfully applied to concrete nonlinear systems and can lead to, for instance, proving the existence of symbolic dynamics in the Lorenz system. See a survey article [7] for details and more information.

This session presents more recent results in this direction that exhibit further developments. The topics include:

• a new topological idea for rigorously proving the existence of various different trajectories of dynamical systems such as periodic, homo/heteroclinic solutions, and symbolic dynamics (P. Zgliczynski)
• a new method for computation of resonant surfaces and Arnold tongues in maps and ODEs (F. Schilder)
• rigorous verification of hyperbolicity of the chain recurrent set of the Hénon map family (Z. Arai)
• computational approach to estimating the lower bound of the measure of parameters with positive Lyapunov exponents in the logistic family (S. Luzzatto)

These results illustrate potential power and future possibilities of computational approach to nonlinear systems for studying their global dynamics and bifurcations. It should also be mentioned that similar ideas can also be applied to systems in infinite dimension, see e.g. [2] and references therein.

Several useful software packages have been developed for studying global dynamics and bifurcations, some of which are used in or related to the above works that are listed below:

AUTO (http://indy.cs.concordia.ca/auto/) Software for continuation and bifurcation problems in ODEs (and maps);

GAIO (http://math-www.uni-paderborn.de/~agdellnitz/Software/gaio.html) Set-oriented computation of global dynamical objects;

CAPD (http://capd.wsb-nlu.edu.pl/) Rigorous numerical computation of dynamical objects;

CHomP (http://www.math.gatech.edu/~chomp/) Computing homology and related objects such as the Betti numbers.

2. Topological horseshoe in a slow-fast system of ODEs

The content of this section is based on a recent joint work with Marcio Gameiro (Georgia Tech., USA), Tomáš Gedeon (Montana State Univ., USA), Bill Kalies (Florida Atlantic Univ., USA), Konstantin Mischaikow (Georgia Tech./Rutgers Univ., USA), and Hiroe Oka (Ryukoku Univ., Japan), and will appear in Journal of Dynamics and Differential Equations (2006).
2.1. Slow-fast system

Dynamical problems with two different time scales are often modeled by singularly perturbed ODEs called slow-fast systems, which exhibit interesting phenomena such as periodic, heteroclinic or chaotic dynamics. Analytic and geometric theories ([5] and references therein) can be applied to slow-fast systems, once certain conditions are verified, such as the basic hypotheses of the fast and slow subsystems and transversality of relevant normally hyperbolic invariant manifolds. However, analysis is not often strong enough to verify these conditions. In contrast, a purely topological theory for slow-fast systems was developed in [4], which enables us to show various dynamical properties of slow-fast systems under weaker hypotheses. Based on this theory, we illustrate how one can rigorously verify the required hypotheses using existing numerical methods, in a particular example studied by Gardner and Smoller[3] on periodic traveling waves in a reaction diffusion system

$$
eq \frac{e^2 u_{xx} + uf(u,v)}{v_{xx} + vg(u,v)} = (1 - u)(u - v),$$

(1)

where $u$ and $v$ are population densities of a prey and a predator, and $\epsilon > 0$ is small. Choosing the traveling wave coordinate $\xi = (x - \theta t)/\epsilon$, it reduces to the slow-fast system

$$\begin{align*}
\dot{u} &= w \\
\dot{w} &= -\theta w - u(1 - u)(u - v) \\
\dot{v} &= \epsilon v \\
\dot{z} &= -\epsilon (\theta z + v(au - b - v))
\end{align*}$$

(2)

where the derivatives are taken with respect to $\xi$. We set $\theta = -0.25$, $a = 1.65$, and $b = 0.25$, in order to illustrate how these new topological techniques can be combined in a straightforward manner with computational techniques to provide rigorous results about the dynamics of slow-fast systems. Our main results are:

**Theorem 2.1** Consider (1) with $a = 1.65$, and $b = 0.25$. For $\epsilon > 0$ sufficiently small there exist two periodic traveling wave solutions with wave speed $\theta = -0.25$ whose profiles with respect to the $v$-variable are indicated in Figure 1.

![Figure 1: Plot of the two primitive periodic solutions to (2) with $\theta = -0.25$, $a = 1.65$, and $b = 0.25.$](image)

**Theorem 2.2** Consider (1) with $a = 1.65$, and $b = 0.25$. For $\epsilon > 0$ sufficiently small there exists a full two-shift of traveling wave solutions with wave speed $\theta = -0.25$. The profiles with respect to the $v$-variable of these solutions are well approximated by arbitrary concatenations of the profiles indicated in Figure 1.

2.2. Fast dynamics

Here we consider the fast system which is obtained by setting $\epsilon = 0$ in (2), namely

$$\dot{u} = w, \quad \dot{w} = -\theta w - u(1 - u)(u - v).$$

(3)

An isolating block for a vector field is a compact domain whose maximal invariant set is contained in its interior and moreover whose boundary has no inner tangency point of the vector field. The goal here is to identify isolating blocks in (3) at different intervals of $v$ that contain two fixed points $(0, 0)$ and $(1, 0)$, in such a way that, in these $v$-intervals, the unstable manifold of the relevant fixed point changes the exit components of the isolating block through a saddle-saddle connection.

![Figure 2: Coarse polygonal decomposition with 2200 vertices and 1308 polygons from 4336 simplices, and its close-ups near the fixed points $(0, 0)$ and $(1, 0)$. The gray shaded region is an isolating block $X$. The black portion of $\partial X$ is the immediate exit set, and red portion is the immediate entrance set.](image)
fields. See [6] for a general description of the construction and properties of such decompositions. For the fast system (3), we first generate a triangulation in the rectangle $[-0.4, 1.4] \times [-0.7, 0.7]$. The polygonal decomposition resulting from this triangulation is shown in Figure 2, which was rigorously computed by checking the transversality of each edge for the entire interval of parameter values $v \in [0.275, 0.7]$ using interval arithmetic. The total time for these computations, including interval arithmetic, was 2.8 minutes on a 3.0 GHz P4 machine.

Once the polygonal decomposition is obtained, we can describe a theoretical foundation for algorithms to compute global qualitative information about the dynamics of flows generated by a system of ODEs. The main idea is to approximate the dynamics on the region $\Omega$ by a multivalued mapping on a polygonal decomposition of $\Omega$. This approximation is determined from the vector field without numerically integrating to approximate specific trajectories. The resulting multivalued map on polygons is a finite, combinatorial representation of the flow from which specific qualitative information can be extracted using topological ideas from the Conley index theory. Notice that the multivalued map for the fast system (3) is valid for all parameter values $v \in [0.275, 0.7]$, since all the edges are transverse for all $v \in [0.275, 0.7]$. Finally the isolating block $X$ was computed starting from the three polygons containing the three fixed points $(0, 0), (0, 3), 0), and $(1, 0)$ as shown in Figure 2. It can be shown that $X$ is an isolating block for all parameter values $v \in [0.275, 0.7]$.

We then used the CAPD Library for computing rigorous enclosures of the unstable manifolds of an equilibrium at different values of $v$. It is clear from Figure 3 that the unstable manifolds connect to different components of the exit set of the isolating block $X$ at the boundary of the corresponding $v$-intervals, and therefore there must exist a saddle-saddle connection in the interior of the intervals. Indeed, we only need a weaker information than the existence of saddle-saddle connections, see [4] for a precise form of the condition.

2.3. Slow dynamics

Now we consider the slow system which is obtained by rescaling $\xi$ and restricting to the slow manifolds defined by $u = 0$ or $u = 1$.

$$v' = z, \quad \xi' = -\theta z - v(au - b - \nu), \quad u = 0 \text{ or } 1. \quad (4)$$

In this case we again used the CAPD Library to obtain bounds for finite time trajectories for several points and check that they satisfy certain geometric constraints.

![Figure 4: Orbits of (4) and its close-ups. The blue & red curves are for $u = 0$ and the green ones for $u = 1$. The vertical black lines are the lines $v = 0.323, v = 0.326$, $v = 0.674$ & $v = 0.677$.](image)

Define two regions between the two black lines given by $v = 0.323$ and $v = 0.326$:

$$\mathcal{R}_R := [0.323, 0.326] \times [0.1, 0.2], \quad \text{and} \quad \mathcal{R}_B := [0.323, 0.326] \times [-0.1, -0.4],$$

and let $\mathcal{R}' := [0.674, 0.677] \times [-0.4, -0.6]$ be a region between the two black lines given by $v = 0.674$ and $v = 0.677$. Note that these regions correspond to the $v$-intervals in which the fast system (3) undergoes the saddle-saddle connections.

Further, we let $\psi_0$ and $\psi_1$ as the flows of the slow system (4) with $u = 0$ and $u = 1$, respectively, and define the sets

- $\mathcal{E}_B$ as a $\psi_0$-flow box starting at some section $\Sigma_B \subset \{v = 0.677\}$ and ending at the black line $\{v = 0.323\}$;
2.4. Existence of periodic solutions and topological horseshoe

Once the sets constructed in the previous subsection are given, the theory developed in [4] guarantees the existence of nice objects in the entire \((u, w, v, z)\)-space, called the periodic corridors, which is a set supposedly containing a periodic orbit. Notice that the unions of nice objects in the entire \((u, w, v, z)\)-space, called the periodic corridor. Letting \(\Pi\) be the projection onto the \((v, z)\)-plane, one can show the existence of periodic corridors for \((2)\) of the form:

\[
\mathbf{PC}_B \subset \mathbf{E}_B \cup \mathbf{B}_B ; \quad \mathbf{PC}_R \subset \mathbf{E}_R \cup \mathbf{B}_R,
\]

where \(B_B; \quad \mathbf{B}_B; \quad \mathbf{B}_R\) are “boxes” with \(\Pi(B_B) = B_i\) and \(\Pi(B_R) = B_i\). See [4] for precise definitions and details of the theory. Using the information obtained in Subsection 2.2, one can indeed prove the following theorem by applying Theorem 1.6 in [4].

**Theorem 2.3** For all sufficiently small \(\epsilon > 0\) there are two periodic solutions \(\Gamma_B\) and \(\Gamma_R\) of the system \((2)\), whose projections to slow variables lie in \(\mathcal{P}_B\) and \(\mathcal{P}_R\), respectively.

A similar idea can be applied to concatenated collections of the periodic corridors \(\mathbf{PC}_B\) and \(\mathbf{PC}_R\) yield periodic orbits of different paths of loops. For instance, \(\mathbf{PC}_B\) followed by \(\mathbf{PC}_R\) implies the existence of a periodic orbit that first goes through the blue and green flow boxes and then through the red and green boxes until it closes up. This may be represented by the symbol \(BR\) for “Blue” followed by “Red.” In the same way, one can prove the existence of infinitely many different periodic orbits corresponding to arbitrarily long words with the symbol \(B\) and \(R\). Furthermore, by taking the Hausdorff limit of increasing sequence of the periodic orbits that are given by a infinite sequence of \(B\) and \(R\), one can show the existence of uncountably many solutions that corresponds to arbitrary sequences of the symbols \(B\) and \(R\).

**Theorem 2.4** For a sufficiently small \(\epsilon > 0\) and any symbolic sequence \(\sigma \in (R, B)^\mathbb{Z}\), the system \((2)\) has a solution

\[
\chi^\sigma(t) = (u^\sigma(t), v^\sigma(t), w^\sigma(t), z^\sigma(t))
\]

which satisfies the following condition: there exists a sequence of intervals \([t_i, t_i^+]\) for \(i \in \mathbb{Z}\) such that, for all \(i \in \mathbb{Z}\),

\[
\Pi(\chi^\sigma([t_i, t_i^+])) \subset E_{\sigma_i} \quad \text{and} \quad \Pi(\chi^\sigma([t_i^+, t_{i+1}])) \subset E_{\sigma_{i+1}}.
\]

**References**


Symbolic dynamics and heteroclinic transitions for the Rossler system

Piotr Zgliczyński†
†Jagiellonian University, Institute of Computer Science, Nawojki 11, 30-072 Kraków, Poland.
Email: zgliczyn@ii.uj.edu.pl

Abstract—For the Rössler system with classical parameter values we give a computer assisted proof of the existence of symbolic dynamics and heteroclinic chains joining period one and period two orbits.

1. Introduction

The Rössler system is given by [6]

\[
\begin{align*}
\dot{x} &= -(y+z) \\
\dot{y} &= x + by \\
\dot{z} &= b + z(x-a)
\end{align*}
\]

where \(a, b \in \mathbb{R}\) are fixed parameters. We chose \(a = 5.7, b = 0.2\), these are parameters values originally considered by Rössler. Numerical simulations strongly suggest that the flow generated by (1) possess a so-called strange attractor. In this note we extend our earlier results from [7], where we proved the existence of an invariant set with complicated dynamics apparently contained in the numerically observed attractor. Here, we show that there are two unstable periodic orbits, \(\gamma_1\), of period one and \(\gamma_2\) of period two and an infinite number of heteroclinic orbits orbits connecting \(\gamma_1\) and \(\gamma_2\) in both directions. The period of the periodic orbits is understood as the period of the Poincaré map on the section \(x = 0\). The proofs are computer assisted, which means that the assumptions of abstract theorems are rigorously verified by computer using interval arithmetics (see [5] for a general introduction to interval methods).

2. Topological tools

The main tool used to obtain an infinite number of periodic orbits is the notion of covering relation [9].

2.1. Covering relations

For the sake of brevity we restrict ourself here the the planar case and maps with one apparently unstable direction. For more general setting see [9].

Definition 1 An \(h\)-set on the plane is a quadruple consisting of

- \(N\) a compact subset of \(\mathbb{R}^2\),
- a homeomorphism \(c_N : \mathbb{R}^2 \to \mathbb{R}^2\), such that
  \[c_N(N) = [-1, 1] \times [-1, 1].\]

We set

\[
\begin{align*}
N^{le} &= \{-1\} \times [-1, 1], \\
N^{re} &= \{1\} \times [-1, 1], \\
S(N)^l &= (-\infty, -1) \times \mathbb{R}, \\
S(N)^r &= (1, \infty) \times \mathbb{R}.
\end{align*}
\]

We define

\[
\begin{align*}
N^{le} &= c_N^{-1}(N^{le}), \\
N^{re} &= c_N^{-1}(N^{re}), \\
S(N)^l &= c_N^{-1}(S(N)^l), \\
S(N)^r &= c_N^{-1}(S(N)^r).
\end{align*}
\]

We will call \(N^{le}, N^{re}, S(N)^l\) and \(S(N)^r\) the left edge, the right edge, the left side and right side of \(N\), respectively.

Hence an \(h\)-set \(N\) on the plane is a product of two intervals in some coordinate system, usually this is a parallelogram.

Definition 2 Let \(N_0, N_1\) be \(h\)-sets on the plane. \(G : N_0 \cup N_1 \to \mathbb{R}^2\), we say \(N_0 G\)-covers \(N_1\) (notation: \(N_0 \overset{G}{\rightarrow} N_1\)) iff

\[G(N_0) \subset S(N_1)^l \cup |N_1| \cup S(N_1)^r\] (2)

and one of the following two conditions hold

\[G(L(N_0)) \subset S(N_1)^l, G(R(N_0)) \subset S(N_1)^r\] (3)
\[G(L(N_0)) \subset S(N_1)^r, G(R(N_0)) \subset S(N_1)^l\] (4)

Condition (2) means that the image of \(N_0\) under \(G\) is contained in the ’horizontal’ strip defined by \(N_1\). Conditions (3) and (4) mean that the ’vertical’ edges of \(N_0\) are mapped to different sides of \(N_1\) (see Fig. 1).

An important, yet trivial, property of the covering relation is its stability under small perturbations. This make it possible to verify it rigorously on computer.

Below we present the main theorem about covering relations.

Theorem 1 [9] Let \(N_i\) for \(i = 0, 1, \ldots, l\) be \(h\)-sets on the plane and \(G_i : N_i \to \mathbb{R}^2\) for \(i = 0, 1, \ldots, l - 1\) be continuous maps. Assume that

\[N_0 \overset{G_0}{\rightarrow} N_1 \overset{G_1}{\rightarrow} N_2 \overset{G_2}{\rightarrow} \cdots \overset{G_{l-2}}{\rightarrow} N_{l-1} \overset{G_{l-1}}{\rightarrow} N_1\] (5)
2.2. Symbolic dynamics, invariant sets

Let $K \in \mathbb{N}$ and $\Sigma_K := \{0, 1, \ldots, K-1\}^\mathbb{Z}$. $\Sigma_K$ is a topological space with the Tichonov topology. On $\Sigma_K$ we have the shift map $\sigma$ given by

$$ (\sigma(c))_i = c_{i+1} $$

Let $A = [\alpha_{ij}]$ be a $K \times K$-matrix,

$$ \alpha_{ij} \in \mathbb{R}_+ \cup \{0\}, \quad i, j = 0, 1, \ldots, K-1. $$

We define $\Sigma_A \subset \Sigma_K$

$$ \Sigma_A := \{c = (c_i)_{i \in \mathbb{Z}} \mid \alpha_{c_0, c_1} > 0\} \quad (6) $$

Obviously, $\Sigma_A$ are invariant under $\sigma$.

Let $Y$ be a topological space, $N \subset Y$ and $f : N \to Y$ be continuous. We define the invariant part of $N$ under $f$ by

$$ \text{Inv}(N, f) = \{x \in N \mid \exists (x_i)_{i \in \mathbb{Z}} \, x_0 = x, \, x_i \in N, \, f(x_i) = x_{i+1}, \, i \in \mathbb{Z}\} $$

2.3. Topological Smale horseshoe

To see full power of Theorem 1 let us consider a topological Smale horseshoe, by which we understand a continuous injective map $G : N_0 \cup N_1 \to \mathbb{R}^2$, where $N_0$, $N_1$ are disjoint h-sets, such that (compare figure 1)

$$ N_i \overset{G}{\rightarrow} N_j , \quad \text{for } i, j = 0, 1. \quad (7) $$

Theorem 2 Let $G$ be a topological horseshoe. There exists $S \subset N_0 \cup N_1$, such that $G(S) = S$ (S is invariant) a continuous surjective map $\pi : S \to \Sigma_2$, such that $\sigma \circ \pi = \pi \circ G$ on $S$. Moreover, for any periodic sequence $c \in \Sigma_2$, the set $\pi^{-1}(c)$ contains a periodic point for $G$ with the same period.

About the proof. It is a direct consequence of Theorem 1. $S = \text{Inv}(N_0 \cup N_1, G)$ and for $x \in S$, $\pi(x) = c_i \in \{0, 1\}$, where $G^i(x) \in N_c_i$.

3. $C^1$ tools. Hyperbolicity

We assume that all maps discussed are $C^1$ injections defined on subsets of the plane.

Observe that the existence of the heteroclinic transition chain

$$ N_0 \overset{G}{\rightarrow} N_0 \overset{G}{\rightarrow} M_1 \overset{G}{\rightarrow} M_2 \cdots \overset{G}{\rightarrow} M_k \overset{G}{\rightarrow} N_1 \overset{G}{\rightarrow} N_1 \quad (8) $$

does not imply that there exist $x_0 \in N_0$, $x_1 \in N_1$ and $y \in N_0$, such that

$$ G(x_0) = x_0, \quad G(x_1) = x_1 $$

$$ G^i(y) \in N_0, \quad \lim_{i \to -\infty} G^i(y) = x_0 $$

$$ G^{k+1+i}(y) \in N_1, \quad \lim_{i \to -\infty} G^i(y) = x_1. $$

For such statement to be true it is required that $\text{Inv}(N_i, G) = \{x_i\}$ for $i = 0, 1$. This can be guaranteed by conditions, the cone conditions, involving derivatives of $G$ on $N_0$ and $N_1$. This issue was discussed in [2]. One of the drawbacks of the approach proposed in [2] was that in the case of the loop $N_0 \overset{G}{\rightarrow} N_1 \overset{G}{\rightarrow} \cdots \overset{G}{\rightarrow} N_{k-1} \overset{G}{\rightarrow} N_0$, the estimates for derivatives of $G^i$ are needed.

3.1. The cone condition

The goal of this section is to introduce a method, which will allow to handle relatively easily the hyperbolic structure on h-sets.

Definition 3 Let $N \subset \mathbb{R}^2$ be an h-set and $Q_N : \mathbb{R}^2 \to \mathbb{R}$ be a quadratic form

$$ Q_N((x, y)) = \alpha x^2 - \beta y^2, \quad (x, y) \in \mathbb{R}^2 \quad (9) $$

where $\alpha, \beta$ are positive constants.

The pair $(N, Q_N)$ we be called an h-set with cones. Sometimes we will drop $Q_N$ and we will say that $N$ is an h-set with cones.

Definition 4 Assume that $(N, Q_N), (M, Q_M)$ are h-sets with cones and let $f : N \to \mathbb{R}^2$ be continuous. Assume that $N \overset{f}{\rightarrow} M$. We say that $f$ satisfies the cone condition (with respect to the pair $(N, M)$) if for any $x_1, x_2 \in N$, $x_1 \neq x_2$ holds

$$ Q_M(f_c(x_1) - f_c(x_2)) > Q_N(x_1 - x_2), \quad (10) $$

460
where \( f_c = c_M \circ f \circ c_N^{-1} \), is the map \( f \) expressed in the coordinates \( c_M \) on \( M \) and \( c_N \) on \( N \).

We have the following theorem.

**Theorem 3** Assume that \( N_i \) for \( i = 0, \ldots, l - 1 \) are h-sets with cones. Let \( f : N_0 \cup \cdots \cup N_{l-1} \rightarrow \mathbb{R}^2 \) be a continuous injective map. Assume that

\[
N_0 \xrightarrow{f} N_1 \xrightarrow{f} N_2 \ldots \xrightarrow{f} N_{l-1} \xrightarrow{f} N_0, \tag{11}
\]

and for \( i = 0, \ldots, l - 1 \) \( f \) satisfies the cone condition with respect to \( N_i \) and \( N_i \mod l \).

Then the following conditions are true

- There exists unique \( x_0 \in N_0 \) such that \( f^i(x_0) \in N_i \mod l \) for \( i = 0, \ldots, l - 1 \).
- If \( f^i(x) \in N_i \mod l \) for all \( i \in \mathbb{N} \), then \( \lim_{i \to -\infty} f^i(x) = x_0 \).
- If \( f^i(x) \in N_i \mod l \) for all \( i \in \mathbb{Z}_- \), then \( \lim_{i \to -\infty} f^i(x) = x_0 \).

About the proof. The existence of \( x_0 \) follows from Theorem 1. The convergence follows from the following observation

\[
L(x) = Q_{N_0}(x - x_0) \text{ is a Liapunov function for } f^k \text{ restricted to the set } \{ x \in N_0, f^i \mod l (x) \in N_i \mod l \}.
\]

### 3.2. How to verify the cone condition?

Assume that \( (N, Q_N) \) and \( (M, Q_M) \) are h-sets with cones and a map \( f : N \to \mathbb{R}^2 \) is \( C^1 \). Observe that (10) cannot be directly verified, because for \( x_1 \to x_2 \) both, the left and the right hand side of (10), converge to the same value. Our intention is to replace (10) with something which is verifiable by computer.

Let \([df_c(N_c)]_l\) be the interval enclosure of \( df_c \) on \( N_c \).

The following lemma was proved in [3]

**Lemma 4** Assume that for any \( B \in [df_c(N_c)]_l \), the quadratic form

\[
V(x) = Q_M(Bx) - Q_N(x) \tag{12}
\]

is positively defined, then for any \( x_1, x_2 \in N_c \) such that \( x_1 \neq x_2 \)

\[
Q_M(f_c(x_1) - f_c(x_2)) > Q_N(x_1 - x_2). \tag{13}
\]

Hence to verify the cone condition it is enough to check whether the interval symmetric matrix

\[
V = [df_c(N_c)]_l^T Q_M [df_c(N_c)] - Q_N \tag{14}
\]

is positively defined.

### 4. Rössler system

In this section we consider the Rössler system (1) with parameter values \( a = 5.7 \) and \( b = 0.2 \) and all presented result refer to this system.

We define two Poincaré sections \( \Theta_+ = \{ (x, y, z) \mid x = 0, x' > 0 \} \) and \( \Theta_- = \{ (x, y, z) \mid x = 0, x' < 0 \} \). Let \( P \) be the Poincaré map on the section \( \Theta_+ \) and \( P_{1/2} \) be a map \( \Theta_+ \) from \( \Theta_- \) and from \( \Theta_- \) to \( \Theta_+ \).

#### 4.1. The existence of two hyperbolic periodic orbits

The following lemma has been rigorously established with computer assistance.

**Lemma 5** There exist three h-sets with cones, such that the following is true

\[
N_0 \cup L_0 \cup L_1 \subset \Theta_+, \tag{15}
\]

\[
N_0 \xrightarrow{P} N_0, \tag{16}
\]

\[
L_0 \xrightarrow{P} L_1 \xrightarrow{P} L_0, \tag{17}
\]

Moreover, \( P \) satisfies cone conditions with respect to pairs \((N_0, N_0), (L_0, L_1)\) and \((L_1, L_0)\).

The h-sets \( N_0, L_0 \) and \( L_1 \) are parallelograms with the sides parallel to approximate unstable and stable directions for \( P \) on \( \Theta_+ \). The centers are approximately given by

\[
\text{center}(N_0) = (0, -8.380942, 2.950007e - 002),
\]

\[
\text{center}(L_0) = (0, -5.424074, 3.108122e - 002),
\]

\[
\text{center}(L_1) = (0, -9.519861, 2.909215e - 002).
\]

The length of the sides are \( 2 \cdot 10^{-4} \).

From Theorems 1 and 3 the following theorem follows.

**Theorem 6** There exists a unique point \( x_0 \in N_0 \), such that \( P(x_0) = x_0 \).

There exists a unique point \( x_1 \in L_0 \), such that \( P(x_1) \in L_1 \) and \( P^2(x_1) = x_1 \).

### 4.2. The existence of heteroclinic chains

To formulate the results about the existence of homoclinic and heteroclinic orbits we need first a few definitions.

**Definition 5** Let \( X \subset \mathbb{R}^2 \) and let \( f : X \to \mathbb{R}^2 \) be a continuous injection. Let \( x_i \in X \) and \( T_i \) for \( i = 0, 1 \) be such \( f^{T_i}(x_i) = x_i \) and for all \( 0 < k < T_i \) holds \( f^k(x_i) \neq x_i \).

For a given \( x \in X \), such that \( f^{T_i}(x) \) is defined for \( i \in \mathbb{Z} \) by \( \gamma_{x} = \{ f^i(x) \mid i \in \mathbb{Z} \} \) we denote a (full) orbit of \( x \).

We say that an orbit \( \gamma = \gamma_x \) is a heteroclinic connection between \( x_i \) and \( x_j \), where \( i \neq j = 0, 1 \) if

\[
\lim_{n \to -\infty} f^{nT_i + k_1}(p) = x_i, \quad \lim_{n \to -\infty} f^{nT_j + k_2}(p) = x_j,
\]

for some \( k_1, k_2 \in \mathbb{N} \).
We say that an orbit $\gamma = \gamma_p$ is homoclinic to $x_i$, where $i = 0, 1$ iff
\[
\lim_{n \to -\infty} f^{nT_i+k_1}(p) = x_i, \quad \lim_{n \to \infty} f^{nT_i+k_2}(p) = x_i.
\]
for some $k_1, k_2 \in \mathbb{N}$.

The following lemma has been rigorously established with computer assistance.

**Lemma 7** There exist h-sets $M_i$, $i = 0, \ldots, 36$, such that $M_i$ are either on the section $\Theta_+$ or $\Theta_-$, such that
\[
L_0 \Rightarrow M_0 \xrightarrow{P_1} M_1 \xrightarrow{P_1} M_2 \xrightarrow{P_1} \cdots M_{36} \xrightarrow{P_1} N_0,
\]
where $P_1$ is either $P$ or $P_1/2$.

Similarly, there exist h-sets $A_i$, $i = 0, \ldots, 27$, such that $A_i$ are either on the section $\Theta_+$ or $\Theta_-$, such that
\[
N_0 \xrightarrow{id} A_0 \xrightarrow{P_2} A_1 \xrightarrow{P_2} A_2 \xrightarrow{P_2} \cdots A_{28} \xrightarrow{P_2} L_0,
\]
where $P_2$ is either $P$ or $P_1/2$.

The h-sets $M_i$ (and $A_j$) has been constructed as follows:

- find an approximate heteroclinic orbits $\gamma$ the periodic points in $N_0$ and $L_0$ (or in the other direction for $A_j$)
- we generate approximatively invariant stable and unstable directions along $\gamma$.
- The h-sets $M_i$ (and $A_j$) are defined as parallelograms with sides parallel to the these invariant directions and with centers at $\gamma$. The length of the sides varied but never was greater than $2 \cdot 10^{-4}$.

**Theorem 8** Each of the following sets of orbits for $P$ contains infinitely many geometrically distinct orbits

- the orbits homoclinic to $x_0$,
- the orbits homoclinic to $x_1$,
- the heteroclinic connections between $x_0$ and $x_1$,
- the heteroclinic connection between $x_1$ and $x_0$.

About the proof: Use Lemma 5 and 7 to construct an infinite number chain of covering relations of the form
\[
N_0 \Rightarrow N_0 \Rightarrow \cdots M_i \cdots A_j \cdots M_i \cdots \Rightarrow N_0 \Rightarrow N_0.
\]

Now we apply Theorem 3 to obtain an infinite number of orbits homoclinic to $x_0$. Proof of the other assertions is analogous.

### 5. Rigorous numerics

The proofs of Lemmas 5 and 7 has been performed with computer assistance. This means that using validated numerics we checked rigorously the conditions in the definition of covering relations and the cone condition.

We used the Lohner algorithm [4] to validate the covering relations and the $C^1$-Lohner [8] for cone conditions. In all computations we used 4-th order Taylor method with fixed time step $h = 0.001$. We used the interval arithmetic from CAPD package [1]. Our program was written in C++, gnu-compiler was used. All computation has been performed, both under Windows and Linux, using Pentium processors 1.7 GHz. The total computation time was below one minute.

### References


Computing Arnol’d tongue scenarios: some recent advances

Frank Schilder

Department of Engineering Mathematics, University of Bristol
University Walk, Bristol BS8 1TR, United Kingdom
Email: f.schilder@bristol.ac.uk

Abstract—Many interesting models in science and engineering involve forced or coupled oscillators. The most striking feature of such systems is the transition between phase locking and quasi-periodicity. Phase locking produces a periodic solution that generically persists under variation of parameters. In contrast, quasi-periodicity is a codimension-one phenomenon, which is thus generically destroyed by perturbation. The manifestation of this transition is a well-known bifurcation diagram called the ‘Arnol’d tongue’ scenario. In this talk we present methods for computing Arnol’d tongue scenarios for maps and ordinary differential equations.

1. Introduction

As a typical example, consider a model of a frequency divider given by the ordinary differential equation (ODE)

\[ \dot{x} + (\beta - \alpha)\dot{x}^3 - (\beta/2 - \alpha)\dot{x} + (1 + \alpha \sin 2t)x = 0. \]  

(1)

The engineering problem is to adjust the free parameters \( \alpha \) and \( \beta \) such that the system’s response is a \( 1:q \) sub-harmonic oscillation with frequency \( 2/q \).

In the zero-forcing case, \( \alpha = 0 \), the dynamics in the phase space \((x, \dot{x}, t) \in \mathbb{R}^2 \times S^1\) decouples into a superposition of two oscillations, a harmonic oscillation in \( S^1 \) with frequency \( \omega_1 = 2 \), and an oscillation in \( \mathbb{R}^2 \) governed by the autonomous ODE

\[ \dot{x} + \beta(\dot{x}^3 - \dot{x}/2) + x = 0. \]

This equation has a family of limit cycles for \( \beta > 0 \) and the frequency \( \omega_2 < 1 \) of these cycles is a strictly monotonically decreasing function of \( \beta \). The product of these two oscillations is a normally attracting two-torus in \( \mathbb{R}^2 \times S^1 \), and its intersection with the hyperplane \( t = 0 \) is an invariant circle of the period-\( \pi \) stroboscopic map of the flow in \( \mathbb{R}^2 \times S^1 \). The time-\( \pi \) map restricted to this circle is a rigid rotation with rotation number \( \varrho = \omega_2/\omega_1 \).

Since the invariant torus is normally hyperbolic, it will persist as a torus for sufficiently small forcing amplitudes \( \alpha > 0 \) [3]. This allows a further reduction of our stroboscopic map in \( \mathbb{R}^2 \) to a circle map that is a perturbation of a rigid rotation. The classification of the dynamics of such a circle map leads to an Arnol’d tongue scenario. It has a countable collection of Arnol’d tongues emanating from points on the zero forcing axis where \( \varrho \) is rational and opening up into the region where forcing is turned on. In between these tongues sit quasi-periodic arcs emerging from all points where \( \varrho \) is irrational.

Fig. 1 shows the Arnol’d tongue scenario of eq. (1) in the \((\beta, \alpha)\) parameter plane. The blue shaded areas are the Arnol’d tongues for \( \varrho \in \{3/7, 2/5, 3/8, 1/3, 3/10, 2/7, 3/11, 1/4\} \), and the red curves the quasi-periodic arcs for \( \varrho = n\sqrt{2}/140, n \in \{41, 38, 35, 31, 29, 27, 25\} \) from left to right, respectively. The system’s response is sub-harmonic with frequency \( \omega_1 \varrho \) for parameter values within an Arnol’d tongue and quasi-periodic for parameter values along the quasi-periodic arcs. Note that the bifurcation diagram in Fig. 1 is not a Strutt-Ince diagram, that is, a stability chart of the zero solution. Fig. 1 shows the existence regions of selected sub-harmonic and quasi-periodic solutions.

In what follows we present algorithms for computing Arnol’d tongue scenarios of maps and ODEs. Our discussion closely follows the presentation in [8, 10]. We consider a unified setting of two-parameter families of maps

\[ x \mapsto f(x, \alpha, \beta), \quad f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^n, \]

and of ordinary differential equations

\[ \dot{x} = f(x, \alpha, \beta), \quad f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}^n, \]

where we assume for simplicity that \( f \) is of class \( C^2 \) in all arguments, or has a higher degree of differentiability if re-
required by a numerical procedure that we employ. Similar to our example, the parameter \( \alpha \) plays the role of a coupling strength and \( \beta \) controls the ratio of frequencies in the system. Both equations are assumed to have a normally hyperbolic invariant torus at \( \alpha = 0 \), that is, an invariant circle in the map setting, and an invariant two-torus in the ODE setting. In both cases the \( (\beta, \alpha) \)-parameter plane exhibits an Arnold tongue scenario.

2. Computation of Arnold tongues

We explain the basic idea of our method with a planar map depending on two parameters \( \alpha \) and \( \beta \) that has a \( \beta \)-family of invariant circles for some \( \alpha > 0 \). Consider this family in the vicinity of a 1/2 Arnold tongue as depicted in Fig. 2 in combined phase and parameter space \( (\beta, x_0, x_2) \). For better orientation we superimposed the \( (\beta, x_1) \)-plane with a sketch of an Arnold tongue scenario in the \( (\beta, \alpha) \)-plane. However, we keep \( \alpha \) constant as indicated with the black line. The \( \beta \)-family of invariant circles forms a tube in \( (\beta, x_0, x_2) \)-space and the intersections with the vertical planes are invariant circles for specific values of \( \alpha \) and \( \beta \). All orbits on the circle are dense for \( \beta \) on a quasi-periodic arc (red). On the other hand, all orbits are asymptotic to a period-two orbit for \( \beta \) inside a 1/2 Arnold tongue (blue shaded). More precisely, the invariant circle contains a stable and an unstable period-two orbit. These orbits collide in period-two saddle-node bifurcations at the boundaries of the tongue (dark blue).

The \( (\alpha, s) \) parametrisation of a resonance surface.

The path traced out by the period-two orbits in combined phase and parameter space is a smooth closed figure-eight-like curve, a circular one-parameter family of period-two orbits. We call this curve a constant-\( \alpha \) cross section. In the paragraphs below we introduce an arc-length parametrisation for such constant-\( \alpha \) cross sections and continue these sections with respect to the parameter \( \alpha \). This means that the sections become closed coordinate curves on a so-called resonance surface; see also [5, 6, 7] and look ahead to Fig. 4. We call this parametrisation of a resonance surface the \( (\alpha, s) \)-parametrisation.

Algorithm for maps. We denote a period-\( q \) orbit with \( x^q := \{x^q_1, x^q_2, \ldots, x^q_{nq}\} \in \mathbb{R}^{nq} \), and call \( \mathbb{R}^{nq} \) the period-\( q \) orbit space. Since any point on a constant-\( \alpha \) cross section is uniquely given by a period-\( q \) orbit \( x^q \) and a value of the parameter \( \beta \), we introduce ordered pairs \( \xi = (x^q, \alpha) \) in combined ‘orbit and parameter space’ \( \mathbb{R}^{nq} \times \mathbb{R} \). We define a distance between two ordered pairs \( \xi = (x^q, \alpha) \) and \( \eta = (y^q, \beta) \) by

\[
d(\xi, \eta) := \left( \frac{1}{q} \sum_{j=1}^{q} \|x^q_j - y^q_j\|_2^2 + (\alpha - \beta)^2 \right)^{\frac{1}{2}},
\]

where \( \|\cdot\|_2 \) denotes the Euclidean norm in \( \mathbb{R}^n \). A set of \( N \) ordered pairs \( \xi_j = (x^q_j, \alpha_j) \) that are equally spaced with respect to the distance (2) along the fundamental domain of a circular family of period-\( q \) orbits satisfies the system

\[
x^q_{1j} = f(x^q_{1j}, \alpha, \beta_j), \tag{3}
\]

\[
x^q_{2j} = f(x^q_{2j}, \alpha, \beta_j), \tag{4}
\]

\[\vdots\]

\[
x^q_{nj} = f(x^q_{nj}, \alpha, \beta_j), \tag{5}
\]

\[
d(\xi_j, \xi_{j+1}) = h, \quad j = 1, \ldots, N - 1, \tag{6}
\]

\[
d(\xi_N, \xi_1) = h, \tag{7}
\]

\[
P(\xi) = 0 \tag{8}
\]

of nonlinear equations, where \( h \) is a constant. \( N \) is the number of the Arnold tongues of an ODE by replacing the period-\( q \) orbit conditions (3)-(5) for maps by a suitable boundary value problem for orbit segments of an ODE.
Solutions of the ‘Arnold tongue’ boundary value problem for different distances. In the top panel we used the distance proposed in [8], which takes only the initial points \( x_i(0) \) at \( t = 0 \) into account. In the bottom panel we used the distance (9) which clearly improves the distribution of orbit segments.

The four-dimensional geometry of the 1/4 resonance surface is depicted in Fig. 4. The coordinate lines in \([8]\) are \((\xi, \eta) : (0, 1) \rightarrow \mathbb{R}^n, i = 1, \ldots, q\) and uses the distance

\[
d(\xi, \eta) := \sqrt{\frac{1}{q} \sum_{i=1}^{q} \int_{t=0}^{1} \| x_i(t) - y_i(t) \|^2 \, dt + (a - b)^2} \quad (9)
\]

between two ordered pairs \( \xi = (x, a) \) and \( \eta = (y, b) \) in combined orbit and parameter space \( \mathcal{X} \times \mathbb{R} \); see [10] for more details.

A graphical illustration of a sample solution of this boundary value problem for example (1) is shown in Fig. 3, which also illustrates an improvement of the method in [8]. The four-dimensional geometry of the 1/4 resonance surface is depicted in Fig. 4. The coordinate lines are \((\beta_j, \alpha_k, x_i^j(0))\), where \( k \) is the continuation step in \( \alpha \)-direction.

The more recent algorithm that we present here outperforms this finite-difference method in all relevant aspects: in computation time, simplicity and accuracy. We first explain this algorithm in the context of maps, which reduces to a method independently developed in [1, 4]; see [9, 10] for more complete overviews.

Algorithm for maps. An invariant circle with irrational rotation number \( \varphi \) of the map \( x \mapsto f(x, \alpha, \beta) \) is a solution of the invariance equation \( u(\theta + 2\pi \varphi) = f(u(\theta), \alpha, \beta) \). That is, the map restricted to the invariant circle \( \{ u(\theta) \mid \theta \in [0, 2\pi] \} \) is conjugate to a rigid rotation with rotation number \( \varphi \). Note that this conjugacy always exists if the map is \( C^2 \) and the rotation number Diophantine.

To compute an approximate solution of the invariance equation we substitute \( u \) with a Fourier polynomial \( u_N(\theta) = \sum_{k=-N}^{N} c_k e^{ik\theta} \) of order \( N \), where \( \theta \) lies on the unit circle and \( j \) denotes the imaginary unit, and compute the coefficients \( c_k \) by collocation on a uniform mesh \( \theta_k = (k + 1)\pi/(N + 1) \), \( k = -N, \ldots, N \). Hence, the Fourier coefficient vectors \( c_k \) are a solution of the finite-dimensional system

\[
\begin{align*}
u_N(\theta_k + 2\pi \varphi) &= f(u_N(\theta_k), \alpha, \beta), \quad (10) \\
P(u_N) &= 0 \quad (11)
\end{align*}
\]

of nonlinear equations, where (11) is again a scalar phase condition. This is a system of \( n(2N + 1) + 1 \) equations for the \( n(2N + 1) \) Fourier-coefficients \( c_k \) and the two parameters \( \alpha \) and \( \beta \). Note that we implicitly exploited that \( c_\varphi = \bar{c}_\varphi \) in this count and that the user-defined rotation number is kept fixed. As before, the system is closed by adding the pseudo arc-length constraint.

Algorithm for ODEs. Similar to the method for Arnold tongues we replace the map in (10) with a boundary value problem for orbit segments on the invariant two-torus.
Fig. 5 shows sample solutions of this boundary value problem along the arc with $q = 41 \sqrt{2}/140$ (leftmost in Fig. 1) for different values of the forcing amplitude $\alpha$. The lines on the torus are orbit segments and the red dots on the invariant circle are the collocation points. As the forcing amplitude increases the distribution of collocation points becomes less and less uniform and the method finally fails. This is also corroborated in Fig. 6, where the estimated interpolation error of our approximation to the invariant circle is shown as a function of the forcing amplitude. We observe that the error seems to shoot up at the same value of $\alpha$ for the three finer discretisations and might conclude that the torus itself breaks up close to this point. However, such a conclusion has to be drawn with caution; see [10].

4. Discussion and outlook

An important feature of the techniques discussed in this paper is, that they enable the use of standard two-point boundary value problem software, such as AUTO [2], for computing Arnol'd tongue scenarios of ODEs. In particular, the new method for computing quasi-periodic arcs is suitable for set up in AUTO.

We made major improvements to the methods proposed in [8], and solved all the problems observed in [8] with the computation of Arnol'd tongues. These improvements include the restriction to the fundamental domain of a constant-$\alpha$ cross section, the new distances, the averaged Poincaré type phase condition and the introduction of a powerful adaptation strategy for the Arnol'd tongue algorithms.

Further directions of development are the detection of bifurcations of periodic orbits on a resonance surface and methods for detecting quasi-periodic bifurcations. Furthermore, there is a need for an algorithm for quasi-periodic tori that includes mesh adaptation, namely, adaptation would help to distinguish between a break-up of a discretisation and a break-up of the computed torus. All ‘obvious’ attempts led to unstable algorithms and this remains an open problem.

Acknowledgements

The author would like to thank Bruce B. Peckham for fruitful discussions and Hinke M. Osinga for her careful reading of and constructive suggestions on a draft of this paper. This work was supported by EPSRC grant GR/R72020/01.

References

Abstract—A rigorous computational algorithm for proving the uniform hyperbolicity of dynamical systems is proposed. Applying the algorithm we prove the existence of many regions of hyperbolic parameters in the parameter plane of the real and complex Hénon family.

1. Introduction

A basic question in dynamical systems is: Is the given dynamical system is structurally stable? By the structural stability theorems (see [11]), this problem is closely related to the problem of uniform hyperbolicity. For example, if the chain recurrent set is uniformly hyperbolic, we know that the system is $\mathcal{R}$-stable, that is, there is no bifurcation in the chain recurrent set.

However, proving hyperbolicity is a difficult problem even for such simple polynomial map as the Hénon map

$$H_{a,b} : \mathbb{R}^2 \to \mathbb{R}^2 : (x,y) \mapsto (a-x^2+by,x) \quad (a,b \in \mathbb{R}).$$

The first mathematical result about the hyperbolicity of the Hénon map was obtained by Devaney and Nitecki [7]. They showed that for any fixed $b$, if $a$ is sufficiently large then the non-wandering set of $H_{a,b}$ is uniformly hyperbolic and conjugate to the full horseshoe map, that is, the shift map of the space of bi-infinite sequences of two symbols.

Later, Davis, MacKay and Sannami [5] observed that besides the uniformly hyperbolic full horseshoe region, there is some parameter regions in which the the Hénon map is uniformly hyperbolic and conjugate to a subshift of finite type. For some parameter intervals of the area preserving Hénon family $H_{a,-1}$, they identified the Markov partitions by describing the configuration of stable and unstable manifolds. Although the mechanism of hyperbolicity at these parameter values is clear by their observations, no mathematical proof of the uniform hyperbolicity has been obtained so far.

The purpose of this talk is to give a brief review of a new general method for proving uniform hyperbolicity of dynamical systems [1]. Applying the method to the Hénon map, we obtain a computer assisted proof of the following theorem, which claims the hyperbolicity of Hénon map on many parameter regions including the regions conjectured by Davis et al.

Theorem 1. There exists a set $P \subset \mathbb{R}^2$ such that if $(a,b) \in P$ then $\mathcal{R}(H_{a,b})$ is uniformly hyperbolic. The set $P$ is illustrated in Figure 1 (colored regions).

Here we denote by $\mathcal{R}(H_{a,b})$ the chain recurrent set of $H_{a,b}$. Since the map is $\mathcal{R}$-stable on $P$, there is no bifurcation in $\mathcal{R}(H_{a,b})$ and hence numerical invariants such as the topological entropy, the number of the periodic points, etc., are constant on each connected component of $P$.

We remark that $P$ is a proper subset of the set of uniformly hyperbolic parameters. We can refine Theorem 1 by performing more computations, which yields a set $P'$ of uniformly hyperbolic parameters such that $P \subset P'$.
The method can also be applied for complex Hénon maps. Figure 2 shows (a subset of) the uniformly hyperbolic parameter region of the family of complex Hénon maps with \( b \) fixed to \(-1\), that is, the family \( H_{a,-1} : \mathbb{C}^2 \to \mathbb{C}^2 \) where \( a \in \mathbb{C} \). This computation is the key step for proving the Hubbard’s conjecture on the hyperbolic horseshoe loci of the complex Hénon maps [2]. Roughly speaking, the conjecture implies that the topology of the “Mandelbrot set” of the Hénon map is quite different from that of the usual quadratic map on the complex plane.

Figure 2: “Mandelbrot” set of the complex Hénon maps

The source codes for the program can be downloaded from the author’s web site

http://www.math.kyoto-u.ac.jp/~arai/

2. Hyperbolicity and Quasi-Hyperbolicity

First we recall the definition of hyperbolicity. Let \( f \) be a diffeomorphism on a manifold \( M \) and \( \Lambda \) a compact invariant set of \( f \). We denote by \( T\Lambda \) the restriction of the tangent bundle \( TM \) to \( \Lambda \).

Definition 2. We say that \( f \) is uniformly hyperbolic on \( \Lambda \), or \( \Lambda \) is a uniformly hyperbolic invariant set of \( f \) if \( T\Lambda \) splits into a direct sum \( T\Lambda = E^s \oplus E^u \) of two \( Tf \)-invariant sub-bundles and there are constants \( c > 0 \) and \( 0 < \lambda < 1 \) such that

\[
\|Tf^n|_{E^s}\| < c\lambda^n \quad \text{and} \quad \|Tf^{-n}|_{E^u}\| < c\lambda^n
\]

hold for all \( n \geq 0 \). Here \( \| \cdot \| \) denotes a metric on \( M \).

We note that this definition involves many ingredients, constants \( c \) and \( \lambda \), a splitting of \( T\Lambda \), and a metric on \( M \). If we try to prove hyperbolicity according to this usual definition, we must control these objects at the same time, and the algorithm would be rather complicated.

To avoid this computational difficulty, we introduce the notion of quasi-hyperbolicity. Recall that the differential of \( f \) induces a dynamical system \( Tf : TM \to TM \). By restricting it to the invariant set \( T\Lambda \), we obtain \( Tf : T\Lambda \to T\Lambda \). An orbit of \( Tf \) is called a trivial orbit if it is contained in the zero section of the bundle \( T\Lambda \).

Definition 3. We say that \( f \) is quasi-hyperbolic on \( \Lambda \) if \( Tf : T\Lambda \to T\Lambda \) has no non-trivial bounded orbit.

This definition is much simpler than that of the uniform hyperbolicity and is a purely topological condition for \( Tf \). It is easy to see that hyperbolicity implies quasi-hyperbolicity. The converse is not true in general, although the hyperbolicity of periodic points and the non-existence of a tangency in \( \Lambda \) follows from the quasi-hyperbolicity.

However, when \( f|\Lambda \) is chain recurrent, these two notions coincide.

Theorem 4 ([4, 12]). Assume that \( f|\Lambda \) is chain recurrent, that is, \( \mathcal{R}(f|\Lambda) = \Lambda \). Then \( f \) is uniformly hyperbolic on \( \Lambda \) if and only if \( f \) is quasi-hyperbolic on it.

Next, we rephrase the definition of quasi-hyperbolicity in terms of isolating neighborhoods. Recall that a compact set \( N \) is an isolating neighborhood (see [10]) with respect to \( f \) if the maximal invariant set

\[
\text{Inv}(f, N) := \{ x \in N : f^n(x) \in N \text{ for all } n \in \mathbb{Z} \}
\]

is contained in \( \text{int} N \), the interior of \( N \). An invariant set \( S \) of \( f \) is said to be isolated if there is an isolating neighborhood \( N \) such that \( \text{Inv}(f, N) = S \).

Note that the linearity of \( Tf \) in the fibers of \( T\Lambda \) implies that if there is a non-trivial bounded orbit of \( Tf : T\Lambda \to T\Lambda \), then its multiplication with a constant is also a non-trivial bounded orbit and hence any compact neighborhood \( N \) of the zero-section of \( T\Lambda \) contains a non-trivial bounded orbit. Therefore, the definition of quasi-hyperbolicity is equivalent to saying that the zero section of the tangent bundle \( T\Lambda \) is an isolated invariant set with respect to \( Tf : T\Lambda \to T\Lambda \).

Furthermore, it suffice to find an isolating neighborhood that contains the zero section.

Proposition 5. Assume that \( N \subset T\Lambda \) is isolated neighborhood with respect to \( Tf : T\Lambda \to T\Lambda \) and \( N \) contains the image of the zero-section of \( T\Lambda \). Then \( \Lambda \) is quasi-hyperbolic.

3. Algorithm

Consider a family of diffeomorphisms \( f_\alpha : \mathbb{R}^n \to \mathbb{R}^n \) that depends on \( r \)-tuple of real parameters \( \alpha = (a_1, \ldots, a_r) \in \mathbb{R}^r \).

Define \( F : \mathbb{R}^n \times \mathbb{R}^r \to \mathbb{R}^n \) and \( TF : T\mathbb{R}^n \times T\mathbb{R}^r \to T\mathbb{R}^n \) by

\[
F(x, \alpha) := f_\alpha(x) \quad \text{and} \quad TF(x, \nu, \alpha) := Tf_\alpha(x, \nu)
\]

where \( x \in \mathbb{R}^n \) and \( \nu \in T_x\mathbb{R}^n \).

We denote by \( \mathbb{F} \) the set of floating point numbers, or, the set of numbers our computer can handle. Let \( \mathbb{F} \) be the set of intervals whose end-points are in \( \mathbb{F} \). Namely,

\[
\mathbb{F} := \{ I = [a, b] \subset \mathbb{R} : a, b \in \mathbb{F} \}.
\]
Similarly, we define a set of $n$-dimensional cubes by
\[ \mathbb{I}^n := \{ I_1 \times \cdots \times I_n \subset \mathbb{R}^n \mid I_i \in \mathbb{I} \} . \]

Let $X, F \in \mathbb{I}^n$ and $A \in \mathbb{I}^n$. We consider these cubes as subspaces of the manifold $M$, the tangent space of $M$ and the parameter space, respectively. What we want to compute is the image of these cubes under the map $F$ and $TF$, namely, $F(X \times A)$ and $TF(X \times V \times A)$. Note that these images are not objects of $\mathbb{I}^n$ nor $\mathbb{I}^n \times \mathbb{R}^n$ in general. By this fact and the effect of rounding errors, we can not hope that a computer can exactly compute these images. Instead, we require that our computer can enclose these images using elements of $\mathbb{I}^n$ and $\mathbb{I}^n \times \mathbb{R}^n$.

Assumption 6. There exists a computational method such that for any cubes $X, V \in \mathbb{I}^n$ and $A \in \mathbb{I}^n$, it can computes cubes $Y \in \mathbb{I}^n$ and $W \in \mathbb{I}^n \times \mathbb{R}^n$ such that $F(X \times A) \subset \text{int} Y$ and $TF(X \times V \times A) \subset \text{int} W$ hold rigorously.

Obviously, if the outer approximations $Y$ and $W$ in Assumption 6 are too large, we can not derive any information of $F$ nor $TF$. As we will mention in the last section, for many classes of dynamical systems including polynomial maps, the rigorous interval arithmetic can be used to satisfy this assumption and gives effectively good outer approximations.

Let $K \subset \mathbb{R}^n$ be a compact set that contains $A$ and $L \subset \mathbb{R}^n$ the product of $K$ and $[-1, 1]^n$. We assume that $K$ is decomposed into a finite union of the elements of $\mathbb{I}^n$, namely,
\[ K = \bigcup_{i=1}^{k} K_i \quad \text{where} \quad K_i \in \mathbb{I}^n . \]

We also decompose the fiber $[-1, 1]^n \subset T_\ast \mathbb{R}^n$ into a finite union of the elements of $\mathbb{I}^n$. By making products of cubes contained in the decompositions of $K$ and $[-1, 1]$, we obtain a decomposition of $L$ such as
\[ L = \bigcup_{j=1}^{\ell} L_j \quad \text{where} \quad L_j \in \mathbb{I}^n \times \mathbb{R}^n . \]

By Assumption 6, we can compute $Y_i \in \mathbb{I}^n$ and $W_j \in \mathbb{I}^n \times \mathbb{R}^n$ such that
\[ F(K_i \times A) \subset \text{int} Y_i \quad \text{and} \quad TF(L_j \times A) \subset \text{int} W_j \]
for any $1 \leq i \leq k$ and $1 \leq j \leq \ell$.

From this information of $Y_i$ and $W_j$, we then construct directed graphs $G(F, K, A)$ and $G(TF, L, A)$ as follows:

- $G(F, K, A)$ has $k$ vertices: $\{v_1, v_2, \ldots, v_k\}$.
- There exists an edge from $v_p$ to $v_q$ if and only if $Y_p \cap K_q \neq \emptyset$.

And similarly,

- $G(TF, L, A)$ has $\ell$ vertices: $\{w_1, w_2, \ldots, w_\ell\}$.
- There exists an edge from $w_p$ to $w_q$ if and only if $W_p \cap N_q \neq \emptyset$.

The most important property of $G(F, K, A)$ is that if there exists $x \in K_p$ that is mapped into $K_q$ by $f_a$ for some $a \in A$, then there must be an edge of $G(F, K, A)$ from $v_p$ to $v_q$. This property also holds for $G(TF, L, A)$.

We use these directed graphs to enclose the chain recurrent set of $f_a$ and the maximal invariant set of $N$. For this purpose, we define the following notions.

Definition 7. Let $G$ be a directed graph. The vertices of $G$, the invariant set of $G$ is defined by
\[ \{v \in G \mid \exists \text{ bi-ininitely long path through } v\} . \]

The vertices of $\text{Scc } G$, the set of strongly connected components of $G$ is
\[ \{v \in G \mid \exists \text{ path from } v \text{ to itself}\} . \]

The edges of these graphs are defined to be the restriction of that of $G$.

Note that by definition, $\text{Scc } G$ is a subgraph of $\text{Inv } G$.

For subgraphs $G$ of $G(F, K, A)$ and $G' \subset G(TF, L, A)$, we define their geometric representations $|G| \subset \mathbb{R}^n$ and $|G'| \subset \mathbb{R}^n \times \mathbb{R}^n$ by
\[ |G| := \bigcup_{v \in G} K_v, \quad \text{or} \quad |G'| := \bigcup_{w \in G'} L_w, \]
respectively.

Proposition 8. For any $a \in A$, $\text{Inv}(f_a, K) \subset |\text{Inv } G(F, K, A)|$ and $\text{Inv}(f_a, L) \subset |\text{Inv } G(TF, L, A)|$ hold. Furthermore, if the inclusion $\mathcal{R}(f_a) \subset \text{int } K$ holds for all $a \in A$, then $\mathcal{R}(f_a) \subset |\text{Scc } G(F, K, A)|$ holds for all $a \in A$.

Now we can describe the algorithm to prove the quasi-hyperbolicity. The algorithm involves the subdivision algorithm [6]. Roughly speaking, this means that if we fails to prove quasi-hyperbolicity, then we subdivide all of the cubes in $K$ and $L$ to have a better approximation of the invariant set, and repeat the whole step until we succeed the proof.

In the following, we first develop an algorithm for a fixed set of parameter values. That is, we fix the set $A$ and try to check if $\mathcal{R}(f_a)$ is quasi-hyperbolic for all $a \in A$. Note that we do not exclude the case $A$ contains only one parameter value, namely, $A = \{a\}$ where $a$ is a $r$-tuple of floating point numbers.

Algorithm 9. (proving quasi-hyperbolicity for $a \in A$)

1. Find $K$ such that $\mathcal{R}(f_a) \subset \text{int } K$ holds for all $a \in A$ and let $L := K \times [-1, 1]^n$.
2. Compute $\text{Scc } G(F, K, A)$ and replace $K$ with $\text{Scc } G(F, K, A)$.
3. Replace \( L \) with \( L \cap (K \times [-1, 1]^n) \).

4. Compute \( \text{Inv} G(TF, L, A) \).

5. If \( |\text{Inv} G(TF, K, A)| \subset K \times \text{int} [-1, 1]^n \) then stop.

6. Otherwise, replace \( L \) with \( |\text{Inv} G(TF, L, A)| \) and refine the decomposition of \( K \) and \( L \) by bisecting all cubes in them. Then goto step 2.

**Theorem 10.** If Algorithm 9 stops, then \( f_a \) is quasi-hyperbolic on \( \mathcal{R}(f_a) \) for every \( a \in A \).

In other words, if \( A \) contains a non-quasi-hyperbolic parameter value, then Algorithm 9 never stops. Therefore, if we want to apply the method for a large family of diffeomorphism, the algorithm should involve an automatic selection of parameter values.

We can also use the subdivision algorithm to realize such a procedure. That is, we will inductively decompose \( A \) into a finite union of the elements of \( \mathcal{IF}^r \) and remove cubes in which the hyperbolicity is proved. We denote by \( \mathcal{A} \) the set of cubes in the decomposition of \( A \) (see [1]).

**References**


Rigorous bounds for the probability of chaos in the logistic family
Stefano Luzzatto
Department of Mathematics, Imperial College,
Exhibition Road, London SW7 2AZ, UK
Email: stefano.luzzatto@imperial.ac.uk

Abstract—We describe some recent results which allow rigorous explicit bounds to be obtained for the measure of the set of “chaotic” parameters in families of one-dimensional maps. These results imply in particular that the measure of the set of stochastic parameters in the quadratic family is at least $10^{-5000}$!

1. Introduction

The quadratic family

$$f_a(x) = x^2 + 2$$

and the (essentially equivalent) logistic family $f_a(x) = ax(1-x)$ are prototypical families of one-dimensional maps and perhaps the simplest mathematical models of biological phenomena, see e.g. [24, 25], which can exhibit a wide variety of dynamical behaviour. These dynamical phenomena can depend very sensitively, and very discontinuously, on the parameter even in very smooth parametrized families, see [26] for a comprehensive survey.

As a background to the results presented in this paper, we give a brief survey of the existing qualitative results in the field.

1.1. Regular and stochastic behaviour

We say that $a \in \Omega$ is a regular parameter if $f_a$ has an attracting periodic orbit; we say that $a \in \Omega$ is a stochastic parameter if $f_a$ admits an ergodic invariant probability measure $\mu$ which is absolutely continuous with respect to Lebesgue and has a positive Lyapunov exponent $\lambda(\mu) = \int \log |f'| \, d\mu$. Then we define

$$\Omega^- = \{a : a \text{ is regular} \} \quad \text{and} \quad \Omega^+ = \{a : a \text{ is stochastic} \}.$$

For families of maps with a single critical point, such as the quadratic and logistic families, it is also known that the sets $\Omega^-$ and $\Omega^+$ are disjoint, see [26]. Moreover, for generic families, including the quadratic and logistic families, with a quadratic critical point both $\Omega^-$ and $\Omega^+$ have positive measure and their union has full measure [3–5, 22]. These are therefore the only two “typical” phenomena.

1.2. Topological structure

The topological structure of the two sets is however very different: $\Omega^-$ is open and dense in $\Omega$ [11, 16, 17, 23] and thus $\Omega^+$, which is contained in the complement of $\Omega^-$, is nowhere dense. The fact that it has positive measure is therefore non-trivial. This was first shown in the ground-breaking work of Jakobson [15] and was generalized, over the years, in several papers; we mention in particular [6, 10, 12, 13, 26, 30, 33, 35, 36] for smooth maps with non-degenerate critical points, [34] for maps with a degenerate (flat) critical point, and [20, 21] for maps with both critical points and singularities with unbounded derivatives.

2. The basic questions

The deep and important results mentioned above are essentially qualitative genericity results, and do not a priori address the following basic questions: Given a particular parameter value $a$

what is the dynamics of $f_a$?

or, failing the possibility of an answer to this question,

what is the probability that $a \in \Omega^+$ or $a \in \Omega^-$?

By probability we mean here the relative measures of $\Omega^-$ and $\Omega^+$.

2.1. Regular parameters

If $f_a$ has a periodic attractor, this is in principle verifiable computationally using a variety of methods. Simó and Tatjer [31, 32] have carried out careful numerical estimates of the length of some of the periodic windows, i.e. the open connected components of $= \Omega^-$ in the relevant region $\Omega$ of parameter space (between the “Feigenbaum” and the “top” parameter values). In this region, using sophisticated numerical techniques, they are able to compute something like half $10^6$ small open intervals belonging to $\Omega^-$ corresponding to periodic attractors of period up to about 24 and of lengths as small as $10^{-30}$. The overall length of the periodic windows they compute is roughly 10% of the total length of $\Omega$ giving

$$|\Omega^-| \geq |\Omega|/10.$$
2.2. Stochastic parameters

A natural question is therefore what happens in the remaining 90% of parameters of $\Omega$? One expects the contribution of the smaller windows to be negligible but of course this may not be the case and in principle it is perfectly possible that 99% of parameters of $\Omega$ belong to $\Omega^+$. A possible strategy is to try to identify or estimate the stochastic parameters. Several conditions of various kinds have been identified which imply that a given parameter value is stochastic [1,8–10,14,27,29] but, apart from some exceptional cases, all these conditions require information about an infinite number of iterations and thus are, to all effects and purposes, uncheckable. It can also be shown that in a formal theoretical sense the set $\Omega^+$ is undecidable, see [2].

It is therefore impossible to answer in general the question of whether a specific parameter value belongs to $\Omega^+$. On the other hand it is possible to make some progress on the question of the measure of $\Omega^+$. Murray [28] gives a heuristic argument in this direction, and we present some rigorous results below.

3. Assumptions

We consider some slightly more general families of one-parameter families of $C^2$ interval maps of the form

$$f_a(x) = f(x) - a$$

for some map $f : I \to I$ with a single quadratic critical point $c$ and the parameter $a$ belonging to some interval $\Omega$. In particular the critical point $c \in I$ does not change with the parameter; this is not an essential condition but simplifies some of the already very technical estimates. For the same reason we shall suppose without loss of generality that the interval $I$ contains strictly the interval $[-1,1]$, that the critical point $c = 0$, and that $f_a(c) > 1$ for all parameters $a \in \Omega$. These conditions always hold up to a linear rescaling which will not affect the argument.

3.1. Computable starting conditions

We start by formulating several conditions in terms of six constants $N, \delta, \iota, C_1, \lambda, \alpha_0, \lambda_0$. First of all we define critical neighbourhoods

$$\Lambda := (-\delta, \delta) \subset (-\delta', \delta') =: \Delta^+.$$ We shall assume without loss of generality that both $\log \delta$ and $\iota \log \delta$ are integers.

(A1) Uniform expansivity outside a critical neighbourhood: for all $a \in \Omega$, $x \in I$, $n \geq 1$ such that $x, f_a(x), \ldots, f_a^{n-1}(x) \not\in \Lambda$ we have:

$$\|f_a^n(x)\| \geq \begin{cases} C_1 e^{\lambda n} & \text{if } f_a^n(x) \in \Delta^+ \\ e^{\lambda n} & \text{if } x \in f_a(\Delta^+) \text{ and/or } f_a^n(x) \in \Lambda. \end{cases}$$

(A2) Random distribution of critical orbits: There exists $\tilde{N} \geq N$ such that $|f_a^n(c)| \geq \delta$ for all $n \leq \tilde{N}$ and

$$|\Omega_{\tilde{N}}| := \|f_a^{\tilde{N}}(c) : a \in \Omega\| \geq \delta.' (1)$$

(A3) Bounded recurrence of the critical orbit: for all $a \in \Omega$ and all $N \geq n \geq 1$ we have

$$|f_a^n(c)| \geq e^{-\alpha_0 n}.$$ (A4) Non-Resonance: There exists an integer $\tilde{N} \geq 1$ such that

$$1 + \sum_{i=1}^{\tilde{N}} \frac{1}{(f_a^i)'(c_0)} \neq 0 \forall k \in [1, \tilde{N}],$$

and

$$1 - \sum_{i=1}^{\tilde{N}} \frac{1}{(f_a^i)'(c_0)} \frac{1}{1 - e^{-\lambda_0 n}} > 0.$$ We emphasize that all of these conditions are computable in the sense that they can be verified in a finite number of steps and using explicit numerical calculations relying only on finite precision and depending only on a finite number of iterations. This does not mean however that their verification in practice is trivial or even easy; we shall discuss below some of the computational issues which arise when applying our results to specific situations.

3.2. Remarks

Condition (A1) says that some uniform expansivity estimates hold outside the critical neighbourhoods, uniformly for all parameter values. This is in some sense the most important condition of all, the general principle being that if we have uniform expansivity on a sufficiently large region of the phase space for all parameter values, then we have nonuniform expansivity on all the phase space for a large region of the parameter space.

Condition (A1) as stated is not a strong assumption per se, but becomes strong if we want it to hold for large $\lambda$, small $\delta$ and/or a large interval of parameter values. On the other hand it is important to have a “large enough” parameter interval relative to the choice of $\delta, \iota$ for otherwise (A1) could be satisfied even though all parameters in $\Omega$ have an attracting periodic orbit (this may happen, for instance, if the attracting periodic orbit always has at least one point in $\Lambda$, then its attracting nature could be invisible to derivative estimates outside $\Lambda$).

The appropriate condition on the size of $\Omega$ is given in (A2) which says that the size of the interval given by the images of the critical points for all parameter values at some time is large enough. Technically this gives a sufficiently “random” distribution of the images to provide the first step of the probabilistic induction argument, showing that critical orbits have small probability of returning close to the critical point. This is, conceptually, the core
of the overall argument and this condition is essentially implicit in all arguments of this kind, in the language of Benedicks-Carleson [6, 7] and other papers which follows similar strategies, such as [20, 21]. Intervals of parameter values satisfying these conditions are called escaping components.

Condition (A3) has been used by Benedicks and Carleson and other people and has proved extremely useful in arguments related to shadowing the critical point. Notice however that this condition only refers to an initial finite number of iterates. It is probably no as essential a condition as (A1) and (A2), but rather more of a technical simplifying assumption, albeit one that is not easy to remove in this context. Binding period arguments analogous to the ones developed here using this condition have been generalized in [8] in order to study the statistical properties of a large class of maps with several critical points.

Condition (A4) is easily checked for parameter neighbourhoods of particularly good parameter values, such as the one which we obtain below. As a first application, we are now ready to state precisely our results. First we define the set in the statement of the Theorem below. We give these conditions explicitly just to give an idea of their general form.

\[ \tau_0 \eta_0 < 1. \]  \hfill (C2)

By classical results all these parameters are stochastic.

\[
\max \left\{ \frac{1}{\tau_1} + \frac{\log(DD_2D_3C_1^{-1}) + 2 \log \log \delta^{-i}}{\log \delta^{-i}} \left( \alpha_1 \tau_1 + \frac{\log(\Gamma_1 D_2 D_3 C_1^{-1} e^{i_1 \tau_1^{-1}}) + 2 \log \log \delta^{-i}}{\tau \log \delta^{-i}} \right) \right\} \leq \gamma_1. \]  \hfill (C3)

\[
\eta := \sum_{j=N}^{\infty} \eta^j = \frac{\eta^N}{1 - \eta} < 1. \]  \hfill (C4)

4. Results

We are now ready to state precisely our results. First we define the set

\[ \Omega^* := \{ a \in \Omega : |(f_a^n)'(c_0)| \geq e^{\lambda n} \forall n \geq 0 \}. \]

By classical results, for any \( a \in \Omega^* \), \( f_a \) admits an ergodic absolutely continuous invariant probability measure and thus \( \Omega^* \subseteq \Omega' \).

**Theorem 1** ([19]). Let \( f_a : I \to I \) be a family of \( C^2 \) unimodal maps with a quadratic critical point, of the form \( f_a(x) = f(x) - a \), for \( a \) belonging to some interval \( \Omega \) of parameter values. Suppose that there exist constants \( N, \delta, \epsilon, C_1, \lambda, \alpha_0, \lambda_0 \) so that conditions (A1)-(A4) and (C1)-(C4) hold. Then

\[ |\Omega'| \geq (1 - \eta)|\Omega|. \]

The spirit of this result is to have a ready-made “formula” for proving rigorously the existence, and for obtaining a lower bound for the probability, of stochastic dynamics in a given family of maps.

A very closely related result was proved by Jakobson in [12], where also explicit conditions on a one-parameter family of maps are formulated, and a lower bound on the set of good parameters, similar to the bound given above, is obtained explicitly in terms of the given assumptions. A direct comparison between the upper estimate and that of [12] is not immediate as the way in which the assumptions are formulated is significantly different; also no application is given there to an explicit family, yielding a concrete bound such as the one which we obtain below. As a first application we let \( \lambda_0 = 0.61 \) and

\[ \Omega^* := \{ a \in \Omega : |(f_a^n)'(c_0)| \geq e^{0.61n} \forall n \geq 0 \} \subseteq \Omega'. \]

By classical results all these parameters are stochastic. Then we have
Theorem 2 ([19]). Let $f_a(x) = x^2 - a$ be the quadratic family for $a \in \Omega := [2 - 10^{-4990}, 2]$. Then

$$|\Omega^-| \geq |\Omega^+| \geq 0.97|\Omega| > 10^{-5000}.$$ 

This is not meant to be in any way “optimal” or “sharp” but rather to show that the explicit constants in the Main Theorem can indeed be calculated explicitly in at least one case. We emphasize that even though the interval of parameter values under consideration here is relatively small, this is the first known rigorous explicit lower bound for the set $\Omega^+$ in any context.

A combination of the numerical methods of Simó and Tatjer with the application of the estimates given here might form a good strategy for obtaining some global bounds for the relative measures of $\Omega^-$ and $\Omega^+$ in the entire parameter interval $\Omega$. Indeed, our methods appear naturally suited to the analysis of small parameter intervals and thus, in some sense, take over precisely where the purely numerical estimates no longer work.

References


[22] Mikhail Lyubich, Almost every real quadratic map is either regular or stochastic, Ann. of Math. (2) 156 (2002), 1–78.


UWB Radio: A Real Chance for Application of Chaotic Communications

Géza Kolumbán† and Tamás Krébesz†

†Department of Measurement and Information Systems, Budapest University of Technology and Economics
H-1521 Budapest Pf. 91, Hungary
Email: kolumban@mit.bme.hu, tomkrebesz@gmail.com

Abstract—For lack of empty radio bands, ultra-wideband (UWB) radio that allows the reuse of already occupied frequency bands is the only solution to accommodate new wireless data communication services. The recovery of the very short and extremely wideband wavelets used in UWB radio is not feasible. This is why noncoherent demodulation techniques have to be used. Introducing a general model for the noncoherent UWB modulation schemes the paper evaluates and compares the noise performances of the noncoherent pulse polarity modulation and the transmitted reference system. Closed-form expressions are given for the noise performance of both modulation schemes. Although the former requires a deterministic carrier, the TR system may use either deterministic or chaotic wavelets.

1. Introduction

Frequency band allocated to UWB radio extends from 3.1 GHz to 10.6 GHz. Since conventional narrowband systems have been operating in that frequency band, the power transmitted in a 1-MHz bandwidth is limited to -41.3 dBm. The bandwidth of transmitted UWB signal must exceed 500 MHz.

Since the recovery of UWB wavelets is not feasible, the following noncoherent UWB modulations schemes are considered here: (i) noncoherent pulse polarity modulation where a frequency-shifted bell-shaped Gaussian wavelet is transmitted, (ii) TR system with deterministic wavelets, and (iii) frequency-modulated chaos-shift keying (FM-DCSK) modulation scheme.

The noise performance of noncoherent chaotic communication schemes lags behind that of coherent conventional systems. Since coherent communications cannot be implemented in UWB application, the advantage of conventional systems over the chaotic ones is lost. Furthermore, in UWB application extremely wideband carriers are required and low cost transceivers featuring extremely low power consumption are needed. These requirements may be fulfilled with chaotic systems. This is why UWB radio offers a real chance for the application of chaotic communications.

2. Modulation schemes

The digital information to be transmitted is mapped to wideband wavelets of very short duration in UWB radio. The wavelets have a fixed waveform in impulse radio [1] and they are chaotic signals in chaotic UWB radio [2]. In the latter, the shape of transmitted wavelets is continuously varying even if the same information bit is transmitted repeatedly.

For the sake of simplicity, only binary systems are considered here. Two classes of modulations exist in UWB radio, namely, one information bit may be mapped to (i) one or (ii) two wavelets.

2.1. Structure of modulated UWB signals

Modulation schemes using one wavelet

The structure of UWB modulations using one wavelet is shown in Fig. 1, where \( g(t) \) denotes the wavelet having an arbitrary waveform, \( T_{ch} \) is the wavelet duration and \( t_{pos} \) is the pulse positioning in a pulse bin \( T_{bin} \). The pulse bin denotes the time elapsed between two consecutive wavelets, its duty is to prevent intersymbol interference in a multipath channel. Pulse positioning, amplitude and polarity of a wavelet may be varied in accordance with the modulation. Since the pulse polarity modulation offers the best noise performance, only the pulse polarity modulation is considered here.

\[ g(t) \]

![Figure 1: Structure of UWB modulation using one wavelet.](image)

Modulation schemes using two wavelets

In this case two wavelets, called chips, are used to transmit one bit information. The first chip serves as a reference, while the second one carries the information.

The structure of modulated signal is shown in Fig. 2, where \( g(t) \) denotes an arbitrary wavelet, \( T_{ch} \) is the chip duration and \( \Delta T \geq T_{ch} \) gives the delay between the reference and the information bearing chips. The best noise performance is achieved by the antipodal modulation scheme, where the information bearing wavelet is equal to the delayed reference one for bit “1,” and to the inverted and delayed reference wavelet for bit “0.” This modulation scheme is frequently referred to as transmitted reference (TR) system.

The unique feature of a TR radio system is that the reference chip is not recovered at the receiver but it is transmitted via the same telecommunication channel as the information bearing chip. This solution makes the TR radio system very robust against the linear and nonlinear channel
distortions, but it has a serious drawback: both the reference and information bearing chips are corrupted by the channel noise. The noisy reference chip results in a noise performance degradation.

The fact that the reference chip is transmitted via the same telecommunication channel is generally considered as a disadvantage, since it is considered only as a loss in transmitted energy per bit. This statement is valid if an AWGN channel is considered. However, the real channels always have distortion, either linear or nonlinear. In case of channel distortion, the modulated carrier has to be correlated with a reference signal distorted in the same manner as the modulated carrier to get the best system performance. A correlation with the original distortion-free reference results in a performance degradation. Since in a TR system both the reference and information bearing chips undergo the same distortion, the TR system offers a better system performance when distortion is present in the channel, provided that the loss caused by the noisy reference chip is less than the gain arising due to the perfect correlation of the reference and information bearing chips. The reference chip should be considered as a test signal used to measure the actual channel characteristics. Consequently, the TR system may be used even in a time-varying channel for data communication.

2.2. Wavelets used in UWB radio

Wavelets with very short duration and ultra-wide bandwidth are used in UWB applications. The bandwidths are 500 MHz and 2 GHz in the narrowband and wideband, respectively, UWB radio systems. Except the maximum value of power spectral density and minimum bandwidth, the UWB radio regulation does not specify the type of UWB wavelets, either a fixed or a chaotic waveform may be used.

UWB radio with deterministic wavelet

Because of its excellent spectral properties, a frequency-shifted bell-shaped Gaussian pulse

\[
g(t) = \sqrt{\frac{2E_b}{k \sqrt{\pi} u_B}} \exp\left(-\frac{t^2}{2u_B^2}\right) \cos(\omega_c t) \tag{1}
\]

is used as deterministic wavelet, where \(k = 1\) and \(2\) when one or two wavelets, respectively, are used to transmit the information, \(\omega_c = \omega_c / 2\pi\) is the center frequency, \(E_b\) gives the energy per bit, and \(u_B\) is determined by the required bandwidth of UWB wavelet [1]. The shapes of deterministic wavelets are shown in the time domain for the bandwidths of 500 MHz and 2 GHz in Fig. 3. Although the duration of wavelets is infinite, their power decreases rapidly as a function of time, consequently, a finite wavelet duration may be defined.

UWB radio with chaotic wavelet

In FM-DCSK, the wavelet \(g(t)\) is a frequency modulated signal. This signal is generated in such a way that a chaotic signal is applied to the modulation input of an FM modulator. Although in the original version of FM-DCSK the chip duration \(T_{ch}\) and delay \(\Delta T\) are identical [3], the reference and information bearing chips may be separated \(\Delta T > T_{ch}\) to prevent intersymbol interference (ISI) in UWB application. The former assures the maximum data rate, while the latter prevents the intersymbol interference caused by the multipath channel.

3. UWB demodulator configurations

The wavelet recovery in UWB radio is not feasible because of the (i) very short pulse duration, (ii) complex wavelet waveform and (iii) channel distortion. This is why a noncoherent demodulation must be used. For the UWB modulation schemes using one and two waveforms, the detection with template signal and the autocorrelation detector, respectively, offer the best system performance. Both of them belong to the class of correlator-based receivers.

3.1. Detection of pulse polarity modulation

In pulse polarity modulation the information is carried by the sign of wavelet. Let \(r_{m}(t) = g(t) + n(t)\) and \(\hat{r}_{m}(t) = \hat{g}(t) + \hat{n}(t)\) denote the received noisy signal before and after channel filtering, respectively. The channel noise described by its sample function \(n(t)\) is modeled by a zero-mean stationary Gaussian process having a uniform two-sided power spectral density of \(N_0/2\). The information \(\hat{b}_m\) may be recovered if \(\hat{r}_{m}(t)\) is correlated by a template signal \(p(t)\) as shown in Figs. 4 and 5. The template signal is a
windowing and weighting pulse
\[ p(t) = \begin{cases} \frac{1}{\sqrt{\tau}}, & \text{if } |t| < \frac{\tau}{2} \\ 0, & \text{otherwise} \end{cases} \] (2)

where \( \tau \) is the observation time period.

Assume that the received and template signals are aligned perfectly as shown in Fig. 5. Then the non-optimum width of template signal is the only source of loss in \( E[z_m] \). Let this loss be expressed by a parameter \( \alpha \)
\[ E[z_m] = \sqrt{\alpha E_b}. \] (5)

The template matching efficiency
\[ e_{tm} = \sqrt{\alpha} = \sqrt{\frac{E[z_m]}{E_b}} \]
is plotted in Fig. 6 against the width of windowing template function, i.e., energy capture time \( \tau \), for two different RF bandwidths and for a 4-GHz center frequency.

Figure 6: Template matching efficiency as a function of the width of template signal. The RF bandwidths of bell-shaped Gaussian impulses are 2 GHz (dashed curve) and 2.5 GHz (solid curve).

Figure 6 shows that the demodulator is very sensitive to the timing error, any timing error reduces the separation of message points in the observation space and, consequently, results in a considerable performance degradation. The template matching efficiency also depends on the RF bandwidth, for 2-GHz bandwidth even the sign of \( E[z_m] \) is inverted in a certain range of energy capture time. Note, this means that the demodulator inverts the received bits.

Figure 6 shows that in optimum case the template detector has about a 3-dB theoretical implementation loss.

The bit error rate of noncoherent pulse polarity modulation built with template detection is obtained from (3) [5]
\[ P_e = \frac{1}{2} \text{erfc} \left( \sqrt{\frac{\alpha E_b}{N_0}} \right) \]

The noise performance of pulse polarity modulation built with template detection is shown in Fig. 7 for different RF bandwidths (BW) and template matching efficiencies. Note that the energy capture time has a strong influence on the noise performance, even worse, Figs. 6 shows that if the energy capture time is about \( 1/f_c \) then the demodulation becomes practically impossible.

3.2. Autocorrelation TR receiver

Due to the special structure of TR signal, the information bits may be recovered from the sign of correlation measured between the reference and information bearing chips.
grator is an integrate-and-dump circuit. That determines the noise bandwidth of receiver. The integrator is an integrate-and-dump circuit. as shown in Fig. 8. The channel filter is a bandpass filter that determines the noise bandwidth of receiver. The integrator is an integrate-and-dump circuit.

\[
r_m(t) = r_m(t) \quad \text{Delay} \Delta T \quad f_r \cdot dt \quad \text{Decision circuit}\n\]

Figure 8: Block diagram of TR autocorrelation receiver.

Both noncoherent UWB impulse radio and FM-DCSK systems belong to the TR systems, but in the former a fixed waveform while in the latter a chaotic carrier is used as \( g(t) \). The advantages of these TR systems are: (i) the optimum decision threshold is always zero, consequently, there is no need for an adaptive threshold control and training sequence, (ii) the reference chip measures the actual channel characteristics, (iii) a simple autocorrelation receiver may be used. The TR systems also have a few disadvantages: (i) transmission of a reference chip results in a loss in the energy per bit \( E_b \) (however, this loss may be reduced if more than one information bearing chips are transmitted after one reference chip), (ii) the reference chip is also corrupted by the channel noise.

Noise performance

The equation developed in [4] for the noise performance of FM-DCSK may be generalized to any kind of TR systems implemented with an autocorrelation receiver. The bit error rate (BER) is obtained as

\[
P_e = \frac{1}{2 \pi \tau} \exp \left( -\frac{E_b}{2N_0} \right) \sum_{j=0}^{2\tau-1} \left( \frac{E_b}{2N_0} \right)^j \sum_{i=1}^{2\tau-1} \frac{1}{2j-i} \left( j + 2Br - 1 \right)
\]

where \( \tau \) denotes the energy capture time of autocorrelation receiver. This equation is valid for both fixed and chaotic wavelets provided that the energy per bit \( E_b \) is kept constant if chaotic wavelets are used.

Note that the noise performance of a TR system depends on the product of \( 2Br \), but the delay \( \Delta T - \tau \geq 0 \) between the reference and information bearing chips has no influence on the noise performance. The dependence on \( 2Br \) reflects the fact that both the reference and information bearing chips are corrupted by the channel noise.

4. Conclusions

The noise performances of UWB impulse radio and chaotic FM-DCSK system have been compared. Figures 7 and 9 show that these systems offer a very similar noise performance.

Acknowledgments

This work was financed by the Hungarian Scientific Research Fund (OTKA) under Grant number T-038083.

References

Ultrawideband Transceiver Platform Based On Chaotic Signals


Institute of Radio engineering and Electronics of RAS,
Mokhovaya st., 11-7, Moscow, Russia, 125009
E-mail: chaos@cplire.ru

Abstract – In this report, application of chaotic signals to ultrawideband (UWB) communications is discussed. This approach allows us to develop effective communication systems for wireless personal area networks (WPAN). Miniature testbed of the developed wireless headset using 3,1–5,1 GHz UWB chaotic signal is described. The obtained results allow us to draw a conclusion that the designed architecture can be used as a universal platform for wireless home applications.

1. Introduction

Since the frequency bandwidth and power spectral mask for an ultrawideband system were released by Federal Communication Commission (FCC) in early 2002 [1], many interesting approaches have been researched for commercial applications in this band of 3.1–10.6 GHz. Especially, low cost and low power consumption become the most important issue as a solution for Wireless Personal Area Network (WPAN). For example, IEEE 802.15.4a standard is pursuing for low data rate, longer range, low cost, low power consumption and location awareness, while IEEE 802.15.3a standard is for high performance and high data rate. The IEEE 802.15.4a standard is currently under development as a low data rate UWB (LDR-UWB) solution. Several kinds of UWB signal sources are proposed such as chaotic signal, impulse, and chirp signal [2].

Wireless UWB communication systems seem to be more attractive technology for «smart home» extending WPAN applications. Besides, numerous supplementary communication devices must also be wireless and compatible, e.g., with mobile phones for better convenience.

In this report architecture of low-power transmitter-receiver platform based on direct chaotic communication technology is introduced for “smart home” systems. Technology restrictions peculiar to UWB communication systems are discussed.

2. Restrictions

UWB wireless networks meet some serious restrictions on emitted signal.

First of all, maximum power spectral density of UWB signals is restricted to ~41.3 dBm/MHz, therefore even for the total band (3.1…10.6 GHz) the emitted power should not exceed ~2.4 dBm.

Second restriction is the peak value of emitted power, which should be less than 0 dBm in 50 MHz band. Consequently, restriction for emitted power in 500 MHz band is 20 dBm and for 2 GHz band is 32 dBm.

Finally, third restriction originates from production technology of electronic devices. As a rule, supply voltage of UWB unlicensed communication systems is less than 2 V, so voltage at the antenna input is no greater than 1 V. If input impedance of antenna is 50 Ohm, then the peak value of emitted power is restricted to 10 mW.

The original idea of UWB communication system was the use of ultra-short pulses as carrier (pulse length of about 150 ps for 7.5 GHz band). According to this approach each information bit corresponds to corresponding ultra-short pulse and this means rather simple structure of receiver.

Some proposals based on such a technology were submitted for IEEE 802.15.3a Standard Committee but it can easily be seen that transmission range is restricted to 10 meters due to restrictions on signal amplitude.

This circumstance was not important for high bit-rate communication standard (IEEE 802.15.3a); however it is important for low bit-rate standard IEEE 802.15.4a, where the transmission range should be 30 m at least. In order to achieve such a transmission range the voltage at the antenna input must be ~3 V, which conflicts with tendency to decrease voltage supply and power consumption. Situation is much more serious if transmission range must be more than 50–60 meters.

As is easy to see the third restriction is more essential for pulse-based communication system. Really, first restriction can be avoided by means of decreasing transmission bit-rate, i.e. decreasing the number of emitted pulses per unit time. Second restriction leaves possibility to transmit data to 1000 m (for 2 GHz band and 1000 bps). But the third restriction reduces maximum transmission range to 10 m.

A solution is encoding information bit by a series of ultra-short pulses. In this case, keeping the energy per one information bit constant it is possible to reduce the energy per one ultra-short pulse proportionally to the number of pulses that encode the bit. It means transition from signals with small spreading factor to signals with large spreading factor.

Signals with large spreading factor are regarded in majority of proposals to IEEE 802.15.4a standard. However the original simplicity of pulse-based communication systems is lost.
3. Chaotic UWB Transceiver

Direct chaotic communication scheme means that chaotic signal is generated and modulated directly in RF or microwave frequency band (see [3–7] for detail).

Chaotic signal has various advantages as UWB source. It has naturally wideband signal that can avoid multipath fading, and can give an inexpensive solution compared to conventional spread spectrum signal. Noncoherent RF transceiver system using chaotic signal can be easily implemented without mixers and PLLs. The entire system is expected to consume only 20 mW power with 0.13-μm CMOS RFIC.

Chaotic UWB transceiver is composed of a chaotic signal generator, a switch modulator, a power amplifier, an envelope detector, a low-noise amplifier, and a digital board as shown in Fig. 1. By using noncoherent detection, the system can be implemented on a simple structure. Signal waveforms at each stage of the transceiver are presented in Fig. 1. UWB logarithmic detector of microwave band is used as envelope detector. It is combined with a low noise amplifier at the detector input and a low-pass filter at the output. Use of noncoherent receiver essentially simplifies the structure of the receiver.

The transceiver operates as follows. The chaotic generator provides continuous noise-like chaotic signal of UWB band of 3.1–5.1 GHz. The output signal of generator is amplified by a factor of about 10 dB, and then is OOK (On-Off Keying) modulated by a switch device. The switch modulator plays two independent roles: one is of Tx OOK modulator, the other is a Tx/Rx switch. In Tx mode, one of the switches connects Tx path from power amplifier to antenna, while the second path from antenna to detector is disconnected. The switch is controlled by information signal from a digital board (plot 2). The modulated signal is described by plot 3, and received signal from an antenna has also the same pattern except including noise and delay spread as is shown in plot 4. The signals are emitted and received with the same antenna. The data stream of pulses is formed (plot 5) at the output of the demodulator and then sampled into digital signal (plot 6) by an analog-to-digital converter (ADC). The data signal is put through to digital board as a sequence of transmitted bits.

The chaotic generator is the most important component of the DCC system. Requirements to generator are stable signal generation, appropriate band-width, continuous oscillation in time domain, etc. In order to implement chaotic generators, several kinds of design approaches were proposed [7].

In this paper, ring-type chaotic generator is used. The circuit consists of three amplifier ICs. These amplifiers are connected in series to form a ring using a frequency-selective coupler. This frequency-selective coupler takes the signal out to output load and feeds a part of signal power at the input of the first amplifier. Therefore, three amplifiers and a coupler make a closed loop (fig. 2).

The chaotic generator was simulated with Agilent’s Advanced Design System (ADS). Analysis of multivariate bifurcation diagram allowed us to choose suitable oscillation mode. Simulation result in time domain is presented in Fig. 3. It looks like non-periodic noise. However, its frequency range is limited. Chaotic signal power is 1 mW in 3.1-5.1 GHz band (Fig. 4).

We use this generator because it produces chaotic signal complying with FCC requirements for the spectral mask of the unlicensed frequency band 3.1-10.6 GHz. Generator design for microwaves was completely accomplished by means of ADS, so it has no low-frequency mathematical model.

To show the difference between low-frequency and microwave models of chaotic generator, we compare them in [7, 8] where simulation results regarding to low-frequency model for three-point oscillator described by the set of equations

\[ \begin{align*}
C_1 \dot{V}_{C1} &= -\alpha_f (-V_{C1}) + I_L \\
C_2 \dot{V}_{C2} &= (1 - \alpha_f) (-V_{C2}) + I_L - I_0 \\
L \dot{I}_L &= -V_{C1} - V_{C2} - R I_L + V_{cc}
\end{align*} \]

with simplest I-V transistor characteristic:

\[ f(V_{C2}) = I_N \cdot \exp(V_{C2}/U_i) \]
and results for the same oscillator simulated in ADS in microwave range are considered. Transistor model (Gummel-Poon model) in ADS is represented by high-order system of ODEs with many phase variables.

Modulator plays two roles: it forms chaotic pulse train according to information bit stream and switches transceiver to receive/transmit mode depending on a signal from digital control board. Modulator is implemented with HMC 336 chip.

Receiver of UWB signals amplifies UWB 100-ns chaotic radio pulses in 3–5 GHz band, demodulates and filters them, then amplifies the envelope signal. Signal amplitude at the receiver output is logarithmic function of signal power at the receiver input. Receiver is assembled on logarithmic detector AD8318. This chip is designed to operate in the range 1 MHz to 8 GHz. Input signal power must be –60 to 0 dBm. Output frequency band is 0 to 60 MHz (maximum frequency of envelope signal).

Digital control board accomplishes data exchange with PC via parallel port, encodes/decodes data, forms data frames, controls modulator transceiver by switching transceiver into transmit or receive mode and, finally, it controls transceiver power consumption.

![Fig. 3. Simulated chaotic signal pattern in time domain (x-axis: 5 nsec/div., y-axis: 1 V/div).](image1)

![Fig. 4. Measured power spectrum of chaotic generator (center frequency = 4 GHz, x-axis: 600 MHz/div., y-axis: 10 dB/div.).](image2)

### 4. Performance evaluation

In order to evaluate the system performance, bit-error-rate (BER) and packet-error-rate (PER) is simulated. First, the system performance is evaluated based on simulations in AWGN channel and various IEEE 802.15.4a multipath channels. IEEE 802.15.4a Channel Model Subcommittee is developing channel model from CM1 to CM9 for LR-UWB environments [9, 10]. CM1 and CM2 are based on line-of-sight (LOS) and non-LOS (NLOS) in indoor residential environments, respectively. CM3 and CM4 are based on LOS and NLOS indoor office environment, respectively, and CM5 and CM6 are based on LOS and NLOS outdoor environments, respectively. Industrial environment is one of the most important environments for IEEE 802.15.4a applications. These environments are represented by CM7 and CM8 for LOS and NLOS. CM9 modeled for rural areas. 100 channel impulse responses are simulated. The BER performance and PER performance for AWGN only, CM1, CM5, CM8 environments are presented in Fig. 5 and Fig. 6, respectively.

When the criteria of the system performance is decided at PER ≤ 10−2, minimum Eb/N0 of about 18.0 – 20.0 dB (depends on channel type) can be expected for the chaotic receivers. Generally, the system experiences little degradation under the multipath environments. These are mainly due to the self-inherent wideband characteristic of the chaotic signal and the large guard band setting in the system design parameter. Furthermore, the envelope detector used at the RX acts as an energy collector which summing up the energy from different paths i.e. similar to the effect of RAKE RX with multiple RAKE fingers.

![Fig. 5. BER performance of the chaotic UWB system in AWGN channel for various types of IEEE 802.15.4a multipath channels (2 GHz bandwidth).](image3)

### 5. Experimental Test-Bed

Transceiver testbeds were investigated in order to:
- give a work-out to transceiver tuning technique;
- check efficiency of the entire communication system;
- evaluate bit-error-ratio;
- demonstrate wireless transmission of real digital streams (digital audio streams);
analyzed the system functionality depending on the distance between transmitter and receiver.

Finally, experimental test-beds of transceiver (Fig. 7) were used to transmit/receive audio bit stream in MP3 format.

6. Conclusions

Considered testbeds comply with requirements of IEEE 802.15.4a standard. The obtained results permit us to conclude that original and flexible UWB platform is designed. It has considerable scalability as with transmission bitrate (100 Kbps to 20–50 Mbps) and with transmission range. These properties give the proposed platform field for different wireless applications, particularly, for «smart home» with bitrates corresponding to 4G requirements, i.e. the discussed UWB transceiver can be seen as a universal platform for wireless home applications.

References

On Waiting Time for Path Reconstruction in Mobile Multi-hop Networks

Keisuke Nakano†, Masakazu Sengoku† and Shoji Shinoda‡

†Dept. of Information Engineering, Niigata University
2-8050, Ikarashi, Niigata 950-2181, Japan.
‡Dept. of Electrical, Electronic and Communication Engineering, Chuo University
Tokyo 112-8551, Japan.
Email: {nakano, sengoku}@ie.niigata-u.ac.jp, shinoda@elect.chuo-u.ac.jp

Abstract—In this paper, we evaluate connectivity of a mobile multi-hop network which connects two fixed nodes S and D assuming that the mobile multi-hop network consists of mobile nodes moving along a street. As a measure of connectivity, we use the mean length of a time interval from the time when S fails to transmit data to D in the absence of a multi-hop path to the time when appearance of a multi-hop path enables S to transmit the data to D. This mean waiting time reflects effects of change of topology with time on connectivity. In another article, we showed approximate formulas to compute the mean waiting time for $x = 2r + \Delta x$, where $x$ is the distance between S and D, $r$ is the wireless communication range and $\Delta x$ is a small positive constant. In this paper, we compute the mean waiting time for $x > 2r + \Delta x$ by computer simulation, evaluate and characterize the mean waiting time from some viewpoints, and show possibility to roughly estimate the mean value for $x > 2r + \Delta x$.

1. Introduction

In mobile multi-hop networks [1], two nodes can be connected through a multi-hop path consisting of mobile nodes even if these two nodes are not directly linked. However, source node S and destination node D do not always have a multi-hop path between them because mobile nodes are randomly distributed in general and communication range is finite. The probability that S and D are instantaneously disconnected can be used to measure how often S and D have no multi-hop path. Denote this probability by $P_{\text{disconnection}}$.

It is often said that mobile multi-hop networking has an advantage of the change of topology with time. Namely, S and D can be reconnected by waiting until a multi-hop path appears between them even if there is no multi-hop path at their communication attempt. For example, in sensor networks, information obtained by a sensor node is carried to a base station through a multi-hop path, and such an information delivery does not always require a strict deadline for arrival of information. In this type of application, change of topology with time may become an advantage because information can be delivered from a sensor node to the base station by waiting for appearance of a multi-hop path even if density of mobile relay nodes is small and $P_{\text{disconnection}}$ is high. In such a case, connectivity should be evaluated by the waiting time from the time when source node fails to send data to the time when a multi-hop path appears between source and destination. Denote by $T_{\text{wait}}$ the waiting time.

In [2], we have derived approximate formulas to compute the mean waiting time in communication between two fixed nodes S and D on a street which are connected by a multi-hop path consisting of mobile nodes moving along the street. In the derivations, just two cases are considered. Case 1 is a case where $x \leq 2r$, where $x$ is the distance between S and D and $r$ is the communication range, and Case 2 is a case where $x = 2r + \Delta x$, where $\Delta x$ is a small and positive constant. Then, in this paper, we compute the mean waiting time for $x > 2r + \Delta x$ by computer simulation, evaluate and characterize the mean waiting time from some viewpoints, and show possibility to roughly but theoretically estimate the mean waiting time for $x > 2r + \Delta x$.

The above evaluations include comparison of data diffusion through a mobile multi-hop network with that by a simple epidemic algorithm.

2. Background and problem statement

In this paper, we consider the same model as used in [2]. Two fixed nodes S and D exist along a street as shown in Fig. 1 (a). Distance between S and D is $x$, and S and D are connected via a multi-hop network consisting of mobile nodes between S and D. Two mobile nodes can be directly linked if the distance between them is not longer than $r$, which is a positive constant. A direct link between a mobile and a fixed node exists if the distance between them is not longer than $r$. Assume that mobile nodes move left towards S or move right towards D at a constant velocity $v$ as depicted in Fig. 1 (b). Assume that mobile nodes moving left and those moving right are distributed according to Poisson Processes with intensity $\frac{1}{4}$, respectively, at the initial moment $t = 0$. Assume that mobile nodes can move independently from other nodes. From these assumptions, the distribution of mobile nodes at time $t > 0$ obeys a Poisson distribution of intensity $\lambda$. Then, we can compute $P_{\text{disconnection}}$ from $\lambda$, $r$ and $x$ by using the theory of random clumping [3].

Define that $T_{\text{on}}$ is length of an ON period which is a time
interval during which S and D have at least a multi-hop path, and $T_{off}$ is length of an OFF period which is a time interval during which S and D do not have any multi-hop paths at all. Although some results on connectivity analysis have been reported as explained in [4], $T_{on}$ and $T_{off}$ were firstly defined in [5] to evaluate connectivity between a mobile and a base station in a street multi-hop cellular network, and characteristics of $T_{on}$ and $T_{off}$ have not been analyzed in detail. In [5], $E[T_{on}]$ is not well characterized although $E[T_{on}]$ is theoretically analyzed in some cases and a simple theoretical model for $E[T_{off}]$ is given.

Under the above assumptions, in [2], $E[T_{off}]$ and $E[T_{off}^2]$ are theoretically computed for $x \leq 2r$ (Case 1) and for $x = 2r + \Delta x$ (Case 2). From these values, the mean waiting time is computed as $E[T_{wait}].$ Unfortunately, however, it is not easy to analyze $E[T_{off}]$ and $E[T_{wait}]$ for $x > 2r + \Delta x$ in the same manner although $E[T_{wait}]$ becomes larger as $x$ increases and, as a result, a long waiting time causes a more serious problem.

With these things as a background, at first, we evaluate the mean waiting time for $x > 2r + \Delta x$ by computer simulation. In a computer simulation, we observe state of the network at time intervals of $\Delta t,$ where $\Delta t = 0.05$ in this paper. Secondly, we try to characterize the mean waiting time for $x > 2r + \Delta x$ by representing the mean waiting time as a function of $P_{disconnection}.$ We show that such a characterization enables us to roughly estimate the mean waiting time for $x > 2r + \Delta x$ especially in the case that the waiting time is long.

Also, we evaluate $E[T_{off}]$ from another viewpoint. Even though mobile nodes between S and D cannot reconstruct a multi-hop path quickly for a small $\lambda$ as mentioned above, these mobile nodes may pass through S and D. Then, if S sends data directly to a mobile node M and the mobile node M sends the data to D after moving toward D, such an epidemic data transmission can deliver data from S to D [6]. If the waiting time in the mobile multi-hop network is long due to small $\lambda,$ the epidemic algorithm may realize shorter waiting time than multi-hop networking. We evaluate a difference in $E[T_{off}]$ between a multi-hop transmission and an epidemic transmission.

3. Mean waiting time for reconstruction of a multi-hop path

In this section, we consider the mean waiting time for reconstruction of a multi-hop path. As mentioned above, we have derived approximate formulas to compute the mean waiting time for $x \leq 2r$ (Case 1) and that for $x = 2r + \Delta x$ (Case 2). We just show numerical results of the formulas with simulation results in Fig. 2. Parameters are selected as follows: $r = 10.0,$ $v = 1.0.$ In the computation of the numerical results, $x = 19.9$ in Case 1 and $x = 20.0 + \Delta x$ in Case 2. In the computer simulation, $x = 19.9$ in Case 1 and $x = 20.1$ in Case 2.

In Fig. 2, the numerical results agree well with the simulation results. From this result, we can confirm validity of the analyses. In this figure, we can observe that $E[T_{wait}]$ for $x \leq 2r$ is quite different from that for $x = 2d + \Delta x,$ although difference in $x$ is very small as explained in [2]. Consider a situation where $x = 2r - \Delta x,$ which is included in Case 1. This situation intuitively seems to be similar to Case 2; however, they are very different in mobile multi-hop networking. In Case 1, communication ranges of S and D have an overlap, and S and D are connected every time a relay node enters the overlap. On the other hand, in Case 2, such an overlap does not exist. As a result, S and D always have to be connected by a multi-hop path formed by at least two mobile nodes. Therefore, in Case 2, even if some relay nodes exist between S and D, connection between S and D depends on relative positions of the mobile nodes. For these reasons, there is a difference in the mean waiting time between Case 1 and Case 2 as shown in Fig. 2.

We try to characterize $E[T_{wait}]$ for $x > 2r + \Delta x$ from some simulation results because theoretical computation of $E[T_{wait}]$ for $x > 2r + \Delta x$ is not easy. For the characterization, we show simulation results for $x = 40.0,$ $60.0$ and $100.0$ in Fig. 3 together with those for $x = 10.01,$ $12.5,$ $15.0,$ $17.5,$ $19.9$ and $20.1,$ which are the mean waiting times in Case 1 and Case 2. From Fig. 3, we can see that $E[T_{wait}]$ in Case 1 are almost the same in this graph, and difference between $E[T_{wait}]$ for $x = 19.9$ and that for $x = 20.1$ is large as mentioned above. For $x > 20.1,$ $E[T_{wait}]$ increases as $x$ increases. It is not easy to estimate $E[T_{wait}]$ for $x > 20.1$ only from such a broad tendency.

Next, to change the viewpoint, we show $E[T_{wait}]$ as a function of $P_{disconnection}$ in Fig. 4. From Fig. 4, we can see that $E[T_{wait}]$ decreases as $x$ increases in Case 1. We can confirm that $E[T_{wait}]$ for $x = 2r + \Delta x$ is much larger than that for $x = 2r - \Delta x$ if $P_{disconnection}$ is large. For $x > 2r + \Delta x,$
However, \( E[T_{\text{wait}}] \) does not change so much even if \( x \) increases. This tendency can be explained as follows: Consider a case where \( x = x_1 \) and another case where \( x = x_2 \), where \( x_1 > x_2 \). For the same \( \lambda \), \( E[T_{\text{wait}}] \) for \( x = x_1 \) is larger than that for \( x = x_2 \). To realize \( P_{\text{disconnection}} = p \), where \( p \) is a constant and \( 0 \leq p \leq 1 \), more density is needed for \( x = x_1 \) than for \( x = x_2 \) as shown in Fig. 5, which shows relations between \( \lambda \) and \( x \) which realize \( P_{\text{disconnection}} = 0.1 \), ..., 0.9. More density decreases the mean waiting time as shown in Fig. 3. It is considered that effects of such opposite factors on \( E[T_{\text{wait}}] \) are balanced for \( x > 2r + \Delta x \) and \( E[T_{\text{wait}}] \) is represented as functions of \( P_{\text{disconnection}} \) close to each other.

From these results, if we represent \( E[T_{\text{wait}}] \) as a function of \( P_{\text{disconnection}} \), we can characterize \( E[T_{\text{wait}}] \) for \( x > 2r + \Delta x \) as follows: For the same \( P_{\text{disconnection}} \), \( E[T_{\text{wait}}] \) for \( x > 2r + \Delta x \) is smaller than that for \( x = r + \Delta x \), is larger than that for \( x = 2r - \Delta x \) if \( P_{\text{disconnection}} \) is large, and is close to that for \( x = 2r + \Delta x \). We can compute \( P_{\text{disconnection}} \) from \( \lambda \), \( r \) and \( x \) theoretically as mentioned above. We have a theoretical formula to compute \( E[T_{\text{wait}}] \) for \( x = 2r + \Delta x \), and can represent \( E[T_{\text{wait}}] \) as a function of \( P_{\text{disconnection}} \) like Fig. 4 from the theoretical computation. From this representation, we can roughly but theoretically estimate \( E[T_{\text{wait}}] \) for \( x > 2r + \Delta x \).

As shown in Fig. 3, \( T_{\text{wait}} \) is very long for a small \( \lambda \). In such a case, it is difficult to construct a multi-hop path between S and D even though mobile nodes pass through S and D during the waiting time. Namely, even a simple method such as an epidemic algorithm described in Section 2 and depicted in Fig. 6 may deliver data from S to D more quickly than multi-hop networking. Figures 7 and 8 show the mean time needed to deliver information from S to D by the epidemic algorithm. Denote the mean time by \( E[T_{\text{wait},e}] \). Then, \( E[T_{\text{wait},e}] = \frac{1}{\lambda} \cdot \frac{\lambda}{2} e^{-\lambda x} \) if \( x \leq 2r \), and \( E[T_{\text{wait},e}] = \frac{1}{\lambda} \left( \frac{1}{2} + x - 2r \right) \) otherwise. In these figures,
the mean waiting time in multi-hop networks is also plotted. For a fair comparison, in multi-hop networks, a waiting time is assumed to be 0 if S and D are connected at a communication attempt, and $E[T_{\text{wait}}|P_{\text{disconnection}}]$ is plotted as the mean waiting time. These results show that the mean delay in information delivery from S to D by multi-hop networking is sometimes longer than that by the epidemic algorithm if $x > 2r$ although multi-hop networking outperforms the epidemic algorithm for $x \leq 2r$. This means that, if $x > 2r$, multi-hop networking is possible to give poorer performance in the mean waiting time than a primitive method like the epidemic algorithm. Multi-hop networking between S and D is no longer suitable to such a situation, and we should use multi-hop networking together with another method like the epidemic algorithm.

4. Conclusions

In this paper, we evaluated connectivity of a mobile multi-hop network which connects two fixed nodes S and D using the mean waiting time for reconnection which is realized by change of topology with time due to mobility of relay nodes. We showed simulation results of the mean waiting time as a function of $P_{\text{disconnection}}$ for $x > 2r + \Delta x$, and showed possibility to roughly but theoretically estimate the mean waiting time for $x > 2r + \Delta x$. We also evaluated the mean waiting time in multi-hop networks through comparison with the mean time to delivery information from S to D using an epidemic algorithm.

References


Abstract— In this paper, we investigate two existing opportunistic wireless spectrum access protocols in wireless networks. In the protocols, multiple heterogeneous unlicensed radio systems will coexist and opportunistically use available spectrum without interfering with licensed users, in order to increase spectrum utilization and simplify spectrum management. We formulate the spectrum access problem and propose a new radio system transmission opportunity (TXOP) based spectrum access control protocol to ensure coexistence of multiple radio systems, which may operate on different frequency bands. We evaluate the performance of the spectrum access protocols on spectrum utilization, fairness in sharing opportunistic spectrum, and scalability. Simulation results show that the proposed spectrum access protocol has high scalability and controllability while maintain high spectrum utilization and fairness.

1. Introduction

Advances in Internet technologies have boosted broadband multimedia services. It is envisioned that broadband multimedia services will also be popular over diverse wireless networks. However, wireless radio spectrum is a limited resource in current generation wireless networks, due to varying national and global policies for allocating frequencies. Traditional spectrum management protocols license fixed frequency bands to operators for a long period. Recent studies have found that more than 70% of spectrum is unutilized in most areas. Efficiently reusing this unused spectrum (also referred to as opportunistic or open spectrum) will promote spectrum sharing and ease the problem of bandwidth limitation [1][2]. Protocols in which spectrum is selected as the need arises have been studied as part of cognitive radio systems using software definable radios to achieve operation in a wide range of frequency bands with high accessibility and utilization. In such protocols, unlicensed users (or secondary users) opportunistically use available spectrum without interfering with licensed users, aiming to increase spectrum utilization and simplify spectrum management.

However, with multiple different type of radio systems expected to share the opportunistic spectrum, a good spectrum access protocol needs to provide fairness across users and minimize communications overhead, while achieving the maximum possible spectrum usage. Two opportunistic spectrum access protocols have been proposed [1]. The first one is based on an analytical model to calculate channel access probabilities for each radio systems to optimize fairness. This protocol is implemented in a centralized manner. It requires the detailed information (e.g., protocols used, number of terminals and traffic statistics in each radio system) to model the protocols and calculate optimal transmission probabilities. Therefore it may achieve the optimal fairness in theory for the specific protocols, but it is impractical and inefficient. The second protocol is proposed to overcome the drawbacks of the first protocol by dynamically adjusting channel access probability with local measurements, instead of calculating the optimal transmission probabilities by analytical model. The second protocol is implemented in a distributed way. High fairness on sharing opportunistic spectrum is observed. However, the second protocol requires that each radio system should accurately measure the channel access durations and has the traffic statistics of other radio systems. It will be technically difficult for a radio system to real time monitor the channel access time of other radio systems. It also requires many rounds operations for the protocol to converge to a stable operation points. This is not desirable in an opportunistic spectrum wireless network, as the available spectrum may dynamically change over time.

In this paper, we evaluate the above two opportunistic wireless spectrum access protocols. We formulate the spectrum access problem and propose a new radio system transmission opportunity (TXOP) based spectrum access control protocol. In this protocol, each radio system will obtain a TXOP by competition or by assignment of a network coordinator. During a beacon period, each radio system will monitor its channel usage. If a radio system uses the channel longer than the TXOP in a Beacon, it will stop transmit packets in the left of the beacon period. As each radio system need only monitor its own channel activity and this can be done by a member of this radio system, it eliminates the requirement of strong inter radio system communications. It also does not require the information of the protocols used for channel access and traffic characteristics. Therefore high scalability and flexibility can be achieved. We simulate and compare the performance of the spectrum access protocols on spectrum utilization, fairness in sharing opportunistic spectrum, and scalability. Simulation results show that the proposed spectrum access proto-
col has high scalability and controllability while maintain high spectrum utilization and fairness performance.

2. Channel Access Model

As we are interested in general opportunistic spectrum access, an abstract channel access model is used as shown in Fig. 1. For simplicity of illustration, we assume two different types of radio systems (A and B) operate in the band. Each radio system has \( N_i, i \in A, B \) active user terminals. It is noted we have worked out on multiple radio systems. The radio system of type A represents a wideband system operating on the frequency channel centered on \( f_2 \). The radio systems of type B represents a narrow band system operating on the three frequency channels (center frequencies \( f_1, f_2, f_3 \)). The type A frequency channel overlaps with all the three type B frequency channels. This model is a simplified version of that used in [1], in which three sets of three narrow band frequency channels are considered, while here only one set of three narrow band frequency channels is studied. A simplified listen before talk (LBT) is used for all radio systems to avoid modeling the detailed protocols [1]. In our paper, we used carrier sense multiple access (CSMA) protocol for channel access, which will be introduced in the next subsection. It is modified later for opportunistic spectrum access. We model the arrival traffic as a Poisson random process with rate \( \lambda_i \) for radio system \( i \), so the interarrival time is negative-exponentially distributed with mean time \( 1/\lambda_i \) s. The radio system access duration is constant with \( T_i \) for radio system \( i \), with \( T_i = \text{Rate}_i/\text{PLen}_i \), where constants \( \text{Rate}_i \) and \( \text{PLen}_i \) are transmission rate and packet length for radio system \( i \). We assume that each radio system always detect its own radio resource allocation, instead of other radio systems assumed in [1].

\[
S = \frac{Ge^{-dG}}{G(1 + 2d) + e^{-dG}}
\]  

where \( G \) is offered traffic load and \( d \) is the normalized propagation delay to packet transmission duration. In the left of the paper, if not specified, we will refer CSMA protocol to nonpersistent CSMA protocol.

3. Opportunistic Spectrum Sharing Protocols

In this section, we will introduce three opportunistic spectrum access protocols. The focus of the protocols is on the coexistence of multiple radio systems, which may operate on different frequency bands. Therefore the most important is to achieve high fairness on spectrum sharing while maintaining high overall spectrum utilization as possible.

3.1. CSMA with Transmission Probability Control

It is obvious that without any transmission control, some radio systems may dominate the spectrum usage due to large traffic and number of terminals in their systems. From the previous introduction of the CSMA protocol, we can find that its channel access behavior can be controlled from three possible ways to differentiate throughput and fairness of spectrum sharing,

- Transmission probability control, which means a transmission probability (denoted by \( \rho_i \)) is used to randomly control the transmission of a packet after a terminal believes it gets the right to transmit the packet over the channel.
- Waiting time control, which randomly controls the duration of waiting after a terminal detects the channel busy. This can be similar to the transmission probability control method. We will not discuss this method in the paper.
• Transmission opportunity control, which is used either deterministically or randomly to control the packet transmission of a terminal/radio system over a period determined by a transmission opportunity. This is in contrast to the random control of each packet in transmission probability based control method. We will discuss more details in the next subsection.

Both protocols proposed in [1] are based on transmission probability control to achieve fairness in either centralized or distributed manner. Although the idea of controlling transmission probability is not new, the authors contributed a general Markov chain model, which can be helpful on performance evaluation and long-term network planning. But as we mentioned in Section I, this centralized optimization model depends on closely on the analytical model and require accurate network information, such as number of stations, identical and specific traffic loads. But the real situation may not match with the desired assumptions. The original analytical model is developed for the simple LBT protocol, but it can be extended to the CSMA protocol as well. Evaluation of the centralized method is left for future work.

3.2. CSMA with Dynamic Transmission Probability

As the above centralized transmission probability based control protocol is heavily affected by the network assumption, it is inefficient and not scalable. It is also difficult to extend the analytical model to multihop networks, which limits its usage in single hop infrastructure wireless networks. Therefore a distributed channel access protocol based on anthropological model is proposed [1]. A homo equalis (HE) society model is used. For simplicity, we call this distributed channel access protocol as CSMA-HE protocol, which uses mainly local information and limited global information.

The CSMA-HE protocol can be simply summarized as below. In the protocol, each radio system learns the transmission probability $p_i$ by itself. Denote $x_i$ as the average cumulative channel access duration for radio system type $i$. With initial $p_i = 1$, each time the probability $p_i$ is updated, as follows [1]:

$$p_i = \max(o, \min(1, p_i + \frac{\alpha_i}{n-1} \sum_{x_i < x_i} x_j - x_i + \frac{\beta_i}{n-1} \sum_{x_i < x_i} x_j - x_i))$$  \hspace{1cm} (2)

where $n$ is the number of different radio system types, and $0 \leq \beta < \alpha_i \leq 1$. The only local information needed is its own and the other radio system’s hisitory of cumulative spectrum access duration.

3.3. CSMA with Transmission Opportunity Control

The idea of controlling channel access by TXOP has been used in IEEE 802.11e standard. TXOP is an important concept introduced for quality of service (QoS) control in 802.11e standard [4]. But it is only used to control individual user terminals. In the proposed protocol, TXOP can be used to control the individual user terminals, but our emphasis is using TXOP to control the channel access of a whole radio system. The protocol is called CSMA-TX protocol.

CSMA-TX protocol can be described as follows. We assume each radio system operates with beacon signals. Beacon signals need not synchronize for different radio systems, and the periods of beacons signals can be different. A terminal/radio system gets a TXOP by competition with other terminals/radio systems or from assignment by access pointer. The TXOP can be fixed over a long time or be changed dynamically according to the network conditions. It is to be used in a predefined beacon period. In the beginning of each new beacon period, a terminal/radio system reset the measured cumulative channel access duration. Then it continuously measure its own cumulative channel access duration. If the access duration of a terminal/radio system exceeds its TXOP before the arrival of the next beacon signal, then the terminal/radio system must partially or fully limit its channel access until the next beacon period.

4. Simulation Results

In this section, we will present typical simulation results on CSMA-P protocol, CSMA-HE protocol, and CSMA-TX protocol. CSMA-P protocol represent the CSMA protocol with neither transmission probability nor TXOP for packet transmission control. The emphasis of the simulations is to verify the feasibility, effectiveness and scalability of the CSMA-HE and CSMA-TX protocols in a single-hop wireless networks. The network scenario is shown in Fig. 2, in which there are 30 wireless terminals located around the wireless access point, with $N_A = 10$, $N_B = 20$. For other pair of terminal numbers, we observed similar results. All the wireless terminals can hear each other and the access point. Each terminal will access the channel and send traffic to the access point by channel access protocols. For simplicity, we assume no packet loss due to low signal-to-noise (SNR). And hidden terminal issues is eliminated by setting propagation delay $d = 0$. Packets are transmitted at 256 kbps. The packet length is fixed as 528 bits.

Fig. 3 presents throughput of radio system A and B, obtained with CSMA-P protocol. It is clear that without external transmission control mechanism, the two radio systems can not fairly share the spectrum. Indeed, the fairness is heavily affected by the traffic load and the number of terminals in each radio system. For example, with $\lambda_A = \lambda_B$, for the normalized traffic load $G = 4.1$, the normalized throughput $S_A$ ($S_B$) of radio system A (B) is 0.42 (1.28). Throughput $S_B$ is 3 times $S_A$; while with the traffic loads of $\lambda_B = 4\lambda_A$, for $G = 4.1$, the normalized throughput $S_A$ ($S_B$) of radio system A (B) is 0.12 (1.38). Throughput $S_B$ is more than 10 times $S_A$.

Fig. 4 presents throughput of radio system A and B, ob-
Figure 2: Network topology. 30 user terminals (\(N_A = 10, N_B = 20\)) randomly located in a 200x200 m area centered at the access point.

Figure 3: Normalized throughput versus normalized traffic load with CSMA-P protocol.

Figure 4: Normalized throughput versus normalized traffic load with CSMA-TXOP protocol.

Figure 5: Normalized throughput versus normalized traffic load with CSMA-HE protocol. \(\lambda_A = \lambda_B\) (solid lines) and \(4\lambda_A = \lambda_B\) (dashed lines). \(\alpha = 0.1, \beta = 0.01\).

5. Conclusion

In this paper, we proposed a new radio system transmission opportunity (TXOP) based spectrum access control protocol, to ensure coexistence of multiple radio systems, which may operate on different frequency bands. We evaluated the performance of the protocol and compared it to the existing spectrum access protocols. Simulation results showed that the proposed spectrum access protocol has high scalability and controllability, while maintaining high spectrum utilization and fairness performance.

References


Investigations on Throughput Gain of MIMO Multiplexing Schemes in HSDPA for High-Speed Mobile Multimedia Services

Hidekazu Taoka, Hiroyuki Aratashi, Kenichi Higuchi, and Mamoru Sawahashi

Radio Access Network Development Department, NTT DoCoMo, Inc.

Abstract—This paper investigates the effect of Multiple-Input Multiple-Output (MIMO) multiplexing employing the Per Antenna Rate Control (PARC) and Dual-stream Transmit Antenna Array (D-TxAA) schemes in High-Speed Downlink Packet Access (HSDPA) based on the system-level simulations assuming urban microcell spatial channel model. Simulation results show that assuming ideal conditions such as unquantized antenna weights and no feedback information error in D-TxAA, D-TxAA increases the user throughput by approximately 3–5% compared to PARC at the 50% cumulative distribution function (CDF) both with Linear Minimum Mean Square Error (LMMSE) and Successive Interference Canceller (SIC) receivers. The results also show, however, that there is no distinct difference observed between the PARC and D-TxAA schemes either in terms of the user throughput or cell throughput employing LMMSE and SIC taking into account a realistic number of feedback information bits and their decoding errors for D-TxAA. Consequently, we show that the user throughput gain employing the 2-by-2 PARC or D-TxAA scheme over a 1-by-2 LMMSE receiver at the 50% CDF is approximately 10% and 17% using the LMMSE and the SIC receiver, respectively.

1. Introduction

High-Speed Downlink Packet Access (HSDPA) was specified as a Release 6 specification in the 3GPP meeting [1]. High-speed mobile multimedia service with the maximum data rate of greater than 14 Mbps will be provided over HSDPA. Moreover, the long-term evolution of the 3G system called Evolved UTRA (UMTS Terrestrial Radio Access) and UTRAN (UMTS Terrestrial Radio Access Network) is under discussion in the 3GPP aiming at a full IP based radio access and radio access network [2]. In parallel, the enhancements to the existing systems, specifications for which were already completed, are necessary such as Wideband-Code Division Multiple Access (W-CDMA), HSDPA, and an enhanced uplink. In particular, the achievement of an achievable user data rate and frequency efficiency higher than the current ones is very important among the system requirements. The Multiple-Input-Multiple-Output (MIMO) transmission technique has recently drawn much attention because of its ability to increase the achievable data rate and to improve the reception quality [3]-[6]. This paper focuses on MIMO multiplexing using spatial division multiplexing (SDM), in which parallel data streams are transmitted from different transmitter antennas. MIMO multiplexing is categorized into the non-pre-coding type employing the sectored beam [7],[8], and the pre-coding type employing a user-specific transmission beam [9]-[11].

This paper presents system-level evaluations on the user throughput and the average cell throughput employing 2-by-2 MIMO multiplexing in HSDPA to clarify the substantial gain achieved by MIMO multiplexing for providing rich mobile multimedia services. In the paper, we assume two MIMO multiplexing schemes: Per-Antenna Rate Control (PARC) belonging to the non-pre-coding type [7] and Dual-stream Transmit Antenna Arrays (D-TxAA) employing pre-coding [9]. Furthermore, we employ a linear minimum mean squared error (LMMSE) receiver [3] or a successive interference cancellation (SIC) receiver [12] in both the PARC and D-TxAA schemes. The rest of the paper is organized as follows. First, Section 2 briefly describes the operations of the PARC and D-TxAA schemes associated with the receiver configurations. Then, Section 3 presents the simulation conditions. Finally, Section 4 presents an evaluation of the gain from MIMO multiplexing in HSDPA and compares the achievable throughput between the PARC and D-TxAA schemes.

2. MIMO Transmission Schemes

In the paper, we assume PARC and D-TxAA as the MIMO multiplexing schemes. Figures 1(a) and 1(b) illustrate the transmitter structures of the PARC and D-TxAA schemes, respectively. In the
signal replica of the branch is subtracted from the composite received
signal. After the removal of the branch with the

- Encoding
- modulation
- Recovered 
data from 
- Tx branch #1
- Recovered 
data from 
- Txbranch #2

GHz). For inter-cell interference modeling, the strongest eight sectors

model (SCM) of the “Urban Microcell” [16] was assumed for the

between sectors are 0.5 and 1.0, respectively. The spatial channel

in interference from non-serving cells [15], which is a favorable condition

It is assumed that the distance-dependent path loss is constant during

with a standard deviation of 10 dB, and instantaneous multipath fading.

sector. However, we set the minimum distance between the Node B

in a wrap around

cell environment. In the system level simulation, we employ a three-

Table 1. System simulation parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transmission power</td>
<td>38 dBm</td>
</tr>
<tr>
<td>Inter-site distance (ISD)</td>
<td>1000 m</td>
</tr>
<tr>
<td>Distance dependent path loss</td>
<td>34.5 + 38log(fc/3)</td>
</tr>
<tr>
<td>Isolation loss</td>
<td>6 dB</td>
</tr>
<tr>
<td>Shadowing standard deviation</td>
<td>10 dB</td>
</tr>
<tr>
<td>Shadowing correlation between cells/sectors</td>
<td>0.5 r 1.0</td>
</tr>
<tr>
<td>Channel model</td>
<td>SCM-Urban Microcell</td>
</tr>
<tr>
<td>UE speed</td>
<td>3 km/h</td>
</tr>
<tr>
<td>Number of antennas</td>
<td>2 (Node-B) / 2 (UE)</td>
</tr>
<tr>
<td>Channel estimation</td>
<td>Ideal</td>
</tr>
<tr>
<td>Transmission time interval (TTI)</td>
<td>2.0 msec</td>
</tr>
<tr>
<td>Hybrid ARQ</td>
<td></td>
</tr>
<tr>
<td>Packet combining scheme</td>
<td>Chase combining</td>
</tr>
<tr>
<td>Round trip delay (RTD)</td>
<td>6TTI (= 12.0 msec)</td>
</tr>
<tr>
<td>Control delay in scheduling and AMC</td>
<td>3TTI (= 6.0 msec)</td>
</tr>
<tr>
<td>Traffic model</td>
<td>Full buffer traffic</td>
</tr>
<tr>
<td>Power allocation to shared data channel</td>
<td>50%, 90%</td>
</tr>
<tr>
<td>Power allocation to common pilot channel</td>
<td>10%</td>
</tr>
<tr>
<td>Spreading factor of shared data channel</td>
<td>16</td>
</tr>
<tr>
<td>Number of multicode for shared data channel</td>
<td>10</td>
</tr>
</tbody>
</table>

3. Simulation Conditions
3.1 Simulation Parameters
Table 1 lists the system-level simulation parameters. In the paper, by
combining a system level simulation and radio link level simulation, we
investigate the user throughput with MIMO multiplexing in a multi-cell
environment. In the system level simulation, we employ a three-
sectored nineteen-hexagonal cell layout model using a wrap around
method to avoid border effects as shown in Fig. 3. The sector antenna
beam pattern has a 70-degree beam width and the inter-site distance is
1000 m. We consider the case of 10 UEs per sector. The locations of
UEs are randomly assigned with a uniform distribution within each
sector. However, we set the minimum distance between the Node B
and a set of UE to 35 meters. The propagation model follows a distance-
dependent path loss with the decay factor of 3.8, lognormal shadowing
with a standard deviation of 10 dB, and instantaneous multipath fading.
It is assumed that the distance-dependent path loss is constant during
the throughput measurement period, while the shadowing and
instantaneous fading variations are added. We assume a 6-dB reduction
in interference from non-serving cells [15], which is a favorable condition
for MIMO deployment owing to the increased probability of a high
SINR region. The correlation values between the cell sites and that
between sectors are 0.5 and 1.0, respectively. The spatial channel
model (SCM) of the “Urban Microcell” [16] was assumed for the
multipath delay profile with the maximum Doppler frequency of fD =
5.55 Hz (corresponding speed of 3 km/h at the carrier frequency of 2
GHz). For inter-cell interference modeling, the strongest eight sectors

are modeled by SCM while the remaining sectors are modeled by
spatially white Gaussian noise processes whose variances are based on
a flat Rayleigh fading process.

The transmission time interval (TTI) is set to 2 msec. The spreading
factor of the High-Speed-Physical Downlink Shared Channel (HS-
PDSCH) is 16, and 10-code channels are used irrespective of the channel
load for the HS-PDSCH. Instead, we change the transmission power
allocation for HS-PDSCH to 50 or 90% according to channel load of
HS-PDSCH (50% transmission power allocation is a typical value
described in [15] and 90% is for reference). Moreover, 10% transmission
power is allocated to the pilot channel and control channels.

We applied adaptive modulation and channel coding (AMC) with
the optimum modulation and coding scheme (MCS) selection threshold
value. Table 2 summarizes the combinations of MCSs. The MCS
updating interval and the AMC control loop delay are set to one TTI
and three TTI, respectively. At the receiver, we assume ideal channel
estimation. Proportional fairness based channel-dependent scheduling
is assumed in each sector with the round trip delay (RTD) of six TTI. We
employ hybrid automatic request (ARQ) with Chase combining with the
RTD of six TTI duration. A full-buffer traffic model is assumed in each sector.

3.2 Link-to-System Interface
The overall procedure of the system level simulation is given hereafter.
1. The SINR value at the output of an LMMSE or SIC receiver is
calculated for each stream per TTI.
2. The receiver feeds back the received SINR of the respective
streams at the output of an LMMSE receiver or an SIC receiver.
We assumed quantized CQI feedback of five bits per stream with the
quantized resolution of 1 dB, CQI feedback error of 0.2 %,
and the CQI feedback delay of three TTI.
3. According to the reported SINR value, channel-dependent
scheduling is applied and the number of transmission streams
and MCSs are selected for the scheduled UE, which satisfies the
initial transmission packet error rate (PER) of less than 10%.
When two streams exist, the packet scheduling is performed
based on the received SINR value averaged over the two streams,
while the MCS selection is independently performed based on
the received SINR value of each stream.
4. Whether or not the packet transmission is successful is determined
by the lookup table of the PER performance in an additive white
Gaussian noise (AWGN) channel using a random draw [9]. When
a packet error occurs, the throughput of the corresponding stream
becomes zero. Otherwise, the throughput value equals the data
rate of the selected MCS.
5. In the case of a packet error, hybrid ARQ is applied. For a retransmitted packet, the received SINR value is updated assuming Chase combining and the packet error is examined using the updated SINR value. Note that the MCS is fixed until the packet is successfully decoded.

In Steps 1 and 4., the calculation of the SINR and throughput is performed assuming the following method.

(1) Calculation method for received SINR and throughput in LMMSE receiver

Weight matrix \( W \) (matrix size of \( N_{TX} (E + \Delta) \times N_{RX} E \)) in an LMMSE receiver is calculated using Eq. (3) [17].

\[
W = H^H (H H^H + \sigma^2 I)^{-1}
\]  

(3)

where \( N_{TX}, N_{RX}, E = (16) \) chips and \( \sigma^2 \) are the interference from other sectors, the number of transmitter antennas, the number of receiver antennas, equalizer window size and cell-sites plus noise power. Furthermore, \( H \) represents the channel matrix (matrix size of \( N_{RX} E \times N_{TX} (E + \Delta) \)) which is given by

\[
H = \begin{bmatrix}
H_{1,1} & \cdots & H_{1,N_{RX}} \\
\vdots & \ddots & \vdots \\
H_{N_{TX},1} & \cdots & H_{N_{TX},N_{RX}}
\end{bmatrix}
\]  

(4)

where \( H_{i,j} \) (matrix size of \( E \times (E + \Delta) \)) consists of the channel impulse response from the \( j \)-th transmitter antenna branch to the \( i \)-th receiver antenna branch of the \( l \)-th path \((l = 1, 2, \ldots, L) \), \( b_i(l) \), and is given by

\[
H_{i,j} = \begin{bmatrix}
h_{i,j}(1) & \cdots & h_{i,j}(L) & 0 & 0 & \cdots & 0 \\
0 & h_{i,j}(1) & \cdots & h_{i,j}(L) & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & \cdots & 0 & h_{i,j}(1) & \cdots & h_{i,j}(L)
\end{bmatrix}
\]  

(5)

In the simulation, we assumed ideal path timing detection and ideal channel estimation, i.e., \( H \), is assumed to be estimated ideally at the UE. It should be noted that this assumption is advantageous for MIMO especially with an SIC receiver, since the impact of the path timing error and channel estimation error strongly affect the signal separation accuracy and the accuracy of the interference cancellation in the SIC. Then, the received SINR of the \( i \)-th stream at the output of an LMMSE receiver is given by

\[
\text{SINR}_{i} = \frac{\|w_i^H (E + \Delta) h_i \|^2}{\sum_j \|w_j^H (E + \Delta) h_j \|^2 + \|w_i^H (E + \Delta) \|^2}
\]  

(6)

where \( w_i \) is the \( j \)-th row vector of weight matrix \( W \) and \( h_i \) is the \( k \)-th column vector of the channel matrix of \( H \). Using Eq. (6), the throughput of each stream is calculated.
receiver achieves almost the same user throughput as PARC employing the SIC receiver. The superior throughput of D-TxAA compared to PARC is caused by the increasing beam gain through pre-coding. The figures also show that with rank adaptation, the user throughput at the 50% CDF employing D-TxAA with ideal transfer of the FBI bits and that with PARC are increased by approximately 20 and 17% compared to that of a single-antenna transmission associated with two-branch antenna diversity reception (1-by-2 LMMSE).

Next, Fig. 6 plots the CDFs of the user throughput using the PARC and D-TxAA receivers assuming realistic conditions for D-TxAA, i.e., with quantized antenna weights based on R99 closed loop mode 1 codebook (two-bit resolution) and 4% FBI error of weight information [14]. Rank adaptation is applied due to its beneficial results in Fig. 5. Similar to the case in Fig. 5, the channel load of HS-PDSCH is set to 50%. Figure 6 shows that the user throughput of D-TxAA with the quantized antenna weights and FBI error is degraded compared to that assuming ideal conditions in Fig. 5 both with the LMMSE and SIC receivers. This is caused by the reducing gain of D-TxAA through quantized antenna weights and the FBI error of 4%. Accordingly, the resultant user throughput is almost the same between D-TxAA and PARC when the LMMSE or SIC receiver is used. Figure 6 also shows that the user throughput gain of 2-by-2 MIMO multiplexing employing PARC or D-TxAA with the LMMSE and SIC receiver over 1-by-2 LMMSE receiver is approximately 10 and 17% at the CDF of 50%, respectively.

Figure 7 plots the corresponding CDFs of the user throughput employing the PARC and D-TxAA receiver when the channel load of HS-PDSCH with MIMO multiplexing is 90%. In this case, all other code channels in the shared data channel with MIMO multiplexing are removed in the SIC. Accordingly, the gain of SIC from LMMSE and that from single-antenna transmission is significantly increased compared to the case with a 50% shared data channel load with MIMO multiplexing. As a result, we see that the user throughput values at the 50% CDF using LMMSE and SIC are increased by approximately 25 and 50% compared to that of the 1-by-2 LMMSE receiver both in the PARC and D-TxAA schemes.

Finally, Table 3 summarizes the sector throughput for the respective MIMO schemes and the gain of MIMO schemes from the single-antenna transmission, i.e., 1-by-2 LMMSE. In this table, the cell throughput gain by 2-by-2 MIMO using PARC or D-TxAA assuming quantized antenna weights with FBI error from the 1-by-2 (3-sector) with a LMMSE receiver is approximately 10% with the LMMSE receiver. Similarly, the corresponding gain is approximately 15–17% with the SIC receiver, respectively.

### Table 3. Average cell throughput and cell throughput gain employing MIMO multiplexing

<table>
<thead>
<tr>
<th>Power allocation to HS-PDSCH</th>
<th>Average cell throughput (Mbps)</th>
<th>Cell throughput gain compared to 1x2 LMMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2x2 PARC (LMMSE)</td>
<td>50% 90% 50% 90%</td>
</tr>
<tr>
<td></td>
<td>2x2 PARC (SIC)</td>
<td>7.11 9.30 1.10 1.24</td>
</tr>
<tr>
<td></td>
<td>2x2 D-TxAA (LMMSE)</td>
<td>7.54 11.20 1.17 1.49</td>
</tr>
<tr>
<td></td>
<td>Quantized weights with FBI error</td>
<td>7.56 11.04 1.17 1.47</td>
</tr>
<tr>
<td></td>
<td>2x2 D-TxAA (SIC)</td>
<td>7.10 9.27 1.10 1.24</td>
</tr>
<tr>
<td></td>
<td>Quantized weights with FBI error</td>
<td>7.74 11.45 1.20 1.52</td>
</tr>
<tr>
<td></td>
<td>1x2 SIMO (LMMSE)</td>
<td>7.44 11.16 1.15 1.48</td>
</tr>
<tr>
<td></td>
<td>Un-quantized weights</td>
<td>6.45 7.52 1.00 1.00</td>
</tr>
</tbody>
</table>

5. Conclusion

This paper investigated the effect of MIMO multiplexing employing the PARC and D-TxAA schemes in HS-DPA based on the system-level simulations assuming an urban microcell spatial channel model. Simulation results showed that assuming ideal conditions such as un-quantized antenna weights and no FBI error in D-TxAA, D-TxAA increases the user throughput by approximately 3–5% compared to PARC at the 50% CDF both with the LMMSE and SIC receivers. The results also showed, however, that there was no distinct difference observed between the PARC and D-TxAA schemes either in terms of the user throughput or cell throughput employing LMMSE and SIC taking into account a realistic number of FBI bits and their decoding errors for D-TxAA. Consequently, it was shown that the user throughput gain of the 2-by-2 PARC or D-TxAA scheme over the 1-by-2 LMMSE receiver at the 50% CDF was approximately 10% and 17% using the LMMSE and the SIC receiver, respectively.

### Reference


[2] 3GPP TR 25.913 (V7.0.0), “Requirements for Evolved UTRA and UTRAN”


494
A CMOS Synapse with STDP Function and Its Application to Hopfield-type Associative Memory

Hideki Tanaka†, Takashi Morie‡ and Kazuyuki Aihara‡

†Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4, Hibikino, Wakamatsu-ku, Kitakyushu 808-0196 Japan
‡Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan, and Aihara Complexity Modelling Project, ERATO, JST
Email: {tanaka-hideki@edu., morie@}brain.kyutech.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—In some spiking neuron models, analog information is expressed by the timing of neuronal spike firing events, and synaptic weights change depending on the relative timing between asynchronous spikes. Therefore, a synapse circuit with STDP (Spike-Timing Dependent synaptic Plasticity) is required for LSI implementation of spiking neural networks with learning function. In this paper, we propose an analog CMOS synapse circuit with symmetric STDP and a Hopfield-type VLSI neural network using the synapse circuit. We have confirmed by circuit simulations that the network can perform autocorrelation learning from input spike patterns.

1. Introduction

A biological neuron receives many electric spike impulses via synapses, and it fires by generating a spike impulse. However, the conventional neural network models usually use analog output values based on the neuronal firing rate coding or the firing population coding. Recently, the spiking neuron models, which express analog information by the timing of neuronal spike firing, attract a lot of attention with expectation of their higher information processing ability [1, 2].

From the theoretical neural network model research, it is expected that more advanced neural systems can be developed using the spiking neuron models. Additionally, since these models operate asynchronously, it is also expected that spiking neural networks operate faster than the conventional synchronous models. So far, the spiking neuron models have mainly been applied to feedforward networks; for example, image data processing using a spiking feedforward network was reported [3].

However, there have been only a few reports about spiking feedback networks. In order to achieve spiking feedback networks with continuous states, spike outputs expressing the zero value are required. However, a neuron generates no spike unless its internal potential exceeds the threshold, and usually the resting (membrane) potential corresponding to the zero value is below the threshold. Therefore, a simple information representation scheme by spike timing cannot be applied to such feedback networks.

We have introduced negative thresholding to solve this problem, and this thresholding operation is achieved by introducing a global excitatory unit (GEU) [4] or by modulating the resting potential [5]. It has also reported that feedback (Hopfield-type) spiking neural networks have a retrieval property for associative memory as conventional Hopfield networks with analog internal potential [4, 5].

On the other hand, it is known that a biological neuron changes its synaptic weights by STDP (Spike-Timing Dependent synaptic Plasticity). STDP is a learning rule depending on relative timing between asynchronous spikes. Therefore, a synapse circuit with STDP is required for LSI implementation of spiking neural networks with learning function.

In this paper, we propose a CMOS synapse with STDP function and a Hopfield-type VLSI neural network using the synapse circuit. There are two types of STDP which are characterized by an asymmetric time window and a symmetric time window [6]. STDP with an asymmetric time window has been realized by analog CMOS circuits and applied to spiking neuron models [7]. The STDP function realized in our circuit is characterized by a symmetric time window. We apply this circuit to the spiking neuron circuit that we have proposed [8] and confirm by circuit simulations that the network can perform autocorrelation learning from input spike patterns.

2. CMOS Circuit for Spiking Neuron Models

The integrate-and-fire-type (IF) neuron model is a typical spiking neuron model. When a spike pulse is fed into a neuron via a synapse, a post-synaptic potential (PSP) is generated. A neuronal internal potential is determined by the spatiotemporal summation of PSPs generated by the input spikes. There are two types of synapses: excitatory and inhibitory, according to the sign of the synaptic weight. When the internal potential exceeds the threshold, a neuron fires and generates a spike. After firing, a refractory period follows. A generated spike is transmitted to other neurons or the neuron itself with a certain transmission delay time. Our spiking neuron circuit, which consists of a synapse
part and a neuron part is shown in Fig. 1. Timing diagrams of the circuit are shown in Fig. 2.

In the synapse part, when \textit{sign} is “High”, the circuit operates as an excitatory synapse, and vice versa. If spike pulse \textit{ii} is fed into the synapse part from other neurons, a PSP control signal \textit{psp\_cont} is generated by a delay-and-inversion circuit \textit{D\&I 2} and a NOR gate. While \textit{psp\_cont} is “High”, a switched-current source turns on and charges or discharges capacitor \textit{C}. Thus, the terminal voltage of the capacitor, \textit{Vn}, is changed, and a PSP is generated. The spatiotemporal summation of PSPs by input spikes is performed by this capacitor. Here, the synaptic weight value is inversely proportional to voltage \textit{Vwij}.

In the neuron part, the internal potential represented by \textit{Vn} returns to resting potential \textit{Vini} by leak resistance connected in parallel with capacitor \textit{C}, after charged or discharged by the current sources of the synapse parts. A comparator (\textit{COMP}) compares \textit{Vn} with threshold voltage \textit{Vth}. If \textit{Vn} exceeds \textit{Vth}, a spike is generated. At the same time, the threshold voltage increases to generate a refractory period. In this circuit, the transmission delay is equal to the refractory period. The \textit{D\&I 1/2} consist of an inverter chain, and their delay time is determined by bias voltages \textit{Vbi} (\(i = 1, 2, \cdots, 6\)).

3. CMOS STDP Circuit

Figure 3 shows a CMOS STDP circuit. The circuit consists of a spike-detection part, a weight-update part and a sign-change part. In the spike-detection part, the state value of a T-FF (Toggle flip-flop) is changed twice by \textit{pre} and \textit{post} spike inputs. Changes in the state value are detected by a \textit{D\&I} and a NOR gate. As a result, a spike pulse which arrives earlier is fed into \textit{in1} of the weight-update part, and the other is fed into \textit{in2}. If spike inputs of \textit{pre} and \textit{post} are given at the same time, the state value of the T-FF changes once. In that case, the T-FF is reset by an AND gate, a \textit{D\&I} and a NOR gate. Depending on only the time difference between spike inputs \textit{in1} and \textit{in2}, the weight-update part updates the synaptic weight represented by \textit{Vwij}. The details are described below.

A timing diagram of the circuit is shown in Fig. 4(a). When the input spike \textit{in1} is fed into the weight-update part from spike-detection part, a ramp signal is generated at the terminal of the capacitor \textit{CA} shown in Fig. 3(b) by the current source. At the same time, control signal \textit{SW}, which is generated by the \textit{D\&I}, turns to “High”. The ramp signal is transformed to a nonlinear waveform by the transistor \textit{M1}, and the waveform is supplied to the input terminal of transconductance-amplifier. The terminal voltage of the capacitor \textit{CB}, \textit{Vb}, returns to referential voltage \textit{VREF} by the resistor \textit{R} after \textit{SW} turns to “Low”. If the input spike \textit{in2} is fed into the circuit during the period of this waveform, the transconductance-amplifier turns on and charges or discharges capacitor \textit{Cwij} to update synaptic weight \textit{Vwij}. If \textit{in2} was given after \textit{Vb} return to \textit{VREF}, the synaptic weight is not updated. A simulation result is shown in Fig. 4(b), which shows a change in the synaptic weight depending on the spike-timing difference.

4. Simulation of Associative Memory

4.1. Simulation Condition

We constructed a Hopfield-type neural network associative memory using the synapse circuit with STDP. The network is composed of 36 neurons with symmetric connections as shown in Fig. 5 [4]. The transmission delay time was set at 200 ns. The simulations were performed by a high-speed circuit simulator, HSIM. In the simulations, the
number of patterns given to memorize was five as shown in Fig. 6. The elements of patterns were randomly chosen under the condition that the numbers of '0' (white pixel) and '1' (black pixel) are equal.

We used spike patterns that expressed the value '1' and '0' with a timing difference of 100ns, and spikes related to the patterns shown in Fig. 6 were given to the network 35 times. If two neurons fire at the same time, the STDP circuit increases the synaptic weight. If the time difference is 100 ns, the synaptic weight are decreased as shown in Fig. 4(b). In the simulations, the maximum values of the increase or decrease were set equal.

After the learning, the STDP function is stopped, and we gave gray-level (5-levels) input patterns that are most similar to pattern #1 shown in Fig. 6 to the network. In the simulations, the time step corresponding to one level in gray-level patterns was set at 25 ns. Therefore, the firing timing of input spikes was limited in \{0, 25, 50, 75, 100\} ns, and the time span for receiving input spikes was 100 ns.

4.2. Simulation Results and Discussion

In the autocorrelation-type associative memory, synaptic weights \( w_{ij} \), which are expressed by the sum of autocorrelation matrices, are generally given by the following equation:

\[
    w_{ij} = \sum_{k=1}^{N} (2I_k^i - 1)(2I_k^j - 1), \text{ for } i \neq j,
\]

where \( w_{ij} = 0 \), and \( I_k^i \) is the \( i \)-th element of the \( k \)-th given pattern vector. In our simulations, because \( N = 5 \), \( w_{ij} \in \{\pm 5, \pm 3, \pm 1\} \). From Eq. 1, the synaptic weights \( w_{12}, w_{46} \) and \( w_{13} \) are 5, 3 and 1, respectively. A simulation result about synaptic weight changes by the STDP function is shown in Fig. 7. In this simulation, the synaptic weights increased approximately by the ratio of 5 : 3 : 1. (It is noticed again that the synaptic weight value is inversely proportional to voltage \( V_{W_{ij}} \)). This result shows that the network can memorize patterns from external spike inputs by using STDP synapses.
5. Conclusions

We proposed a CMOS synapse circuit with symmetric STDP function, and applied to the spiking neuron models. In addition, we constructed a Hopfield-type neural network using the synapse circuit and confirmed by circuit simulations that the network can memorize input spike patterns using STDP function.

Acknowledgment

This work was partially supported by a COE program (center #J19) granted to Kyushu Institute of Technology by MEXT of Japan.

References


[4] K. Sasaki, T. Morie, and A. Iwata, “A Spiking Neural Network with Negative Thresholding and Its Application to Associative Memory,” in IEEE Int. Mid-
Abstract—We discuss the effectiveness of the physical chaotic neuro-dynamics on combinatorial optimization problems. We take two different approaches. The first approach uses the chaotic neuro-dynamics in a framework of the artificial neural networks. In the second approach, the chaotic neuro-dynamics drives a heuristic algorithm, which is a tabu search algorithm. We solve the quadratic assignment problems as a benchmark, and compare the performance and the dynamics of both approaches. Throughout the experiments, we use specially designed hardware systems with analog chaotic neuron integrated circuits, which produce physical chaotic neuro-dynamics for chaotic solution search.

1. Introduction

The chaotic search through a chaotic itinerancy dynamics [1] is known to be very efficient in combinatorial optimization problems [2, 3]. In particular, chaotic dynamics has been used to overcome the local minima problem [4]–[12]. Furthermore, chaotic dynamics not only solves the local minima problem, but also searches the solution more efficiently than stochastic noise under some situations [2, 13]–[15].

Most of the meta-heuristic approaches using chaotic dynamics have been investigated through digital computer simulations. However, these approaches still require enormous computational power and resources for a large-scale optimization problem. Therefore, we require a dedicated high-speed hardware system for real-world combinatorial optimization problems.

The efficiency of the chaotic search comes from the complexity of real-number in chaotic dynamics. A digital computer cannot handle almost all real-numbers, so that simulations with a digital computer would not fully exploit the chaotic search dynamics. On the contrary, an analog circuitry can handle real-numbers in a sense that its state variables are continuous. Therefore, physical chaos from an analog circuit is a good candidate for implementation of the chaotic search [16]. We should underscore here that we must consider the existence of noise in any analog circuit.

Taking the above considerations into account, we have built two mixed analog/digital hardware systems based on the chaotic neural network model [3]. The first system is a chaotic neuro-computer system with switched-capacitor chaotic neurons [17, 18]. The second one is an exponential chaotic tabu search hardware with switched-current chaotic neuron integrated circuits [19].

In this paper, we show several results from these physical chaotic systems. In particular, we will concentrate on the results for the quadratic assignment problems. The quadratic assignment problem (QAP) covers many real-world problems such as a facility layout problem, a routing problem, and a scheduling problem [20]; therefore, the QAP is especially important for engineering applications. However, the QAP is a typical $NP$-hard problem, so that it would take an impractical time to obtain the optimal solution. Consequently, heuristic and meta-heuristic approaches including chaotic search, which find good near optimum solutions in a reasonable time, are very important in real situations [20].

2. Quadratic Assignment Problem

A QAP of size $n$ has two $n \times n$ matrices. One matrix $A$ gives “distances” between arbitrary two “locations.” The other matrix $B$ defines “flows” between arbitrary two “units.” An assignment of the units to the locations are expressed by a permutation $p$ as

$$p : (p(1), p(2), \ldots, p(n))$$

where the elements of $p$ show the units assigned to the locations given by the corresponding indices of $p$.

We should find $p$ which gives the minimum of the cost function $FP$ given by

$$FP = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} b_{p(i)p(j)},$$

where $p(i)$ is the $i$th element of $p$, $a_{ij}$ is the $(i, j)$th element of $A$, and $b_{p(i)p(j)}$ is the $(p(i), p(j))$th element of $B$.

3. QAP with Chaotic Neuro-Computer System

The chaotic neuro-computer system is shown in Fig. 1. The system consists of switched-capacitor chaotic neuron
ICs and digital synapse ICs [17, 18]. We construct 10 × 10 chaotic Hopfield-type neural network for size-10 QAPs in the chaotic neuro-computer. The standard energy function formulation is used to determine the connection weights between neurons.

Because the QAP is in a very difficult class of \( NP \)-hard problem, the bare Hopfield network hardly solves the QAP. Actually, it is difficult to obtain even a feasible solution from the Hopfield-type network. Therefore, we adopt a solution constructing method proposed in [21]. As a result, the optimal solution for Tai10a benchmark problem [20] has been experimentally obtained with the hardware system shown in Fig. 1 [22].

Examples for the time evolutions of the best solution and the energy function are shown in Fig. 2 where the system started the chaotic search at around 120th iteration and finally found the optimal solution at 777th iteration [22].

Figure 1: An external view of the chaotic neuro-computer hardware system.

Figure 2: Examples for the time evolutions of the minimal error from the optimal solution and energy function obtained from the chaotic neuro-computer hardware.

### 4. QAP with Exponential Chaotic Tabu Search Hardware

Yet another approach to solve the QAP is a combination of tabu search algorithm and chaotic neuro-dynamics, i.e., the exponential chaotic tabu search [5]. We have constructed a hardware system for the exponential chaotic tabu search technique using switched-current chaotic neuron ICs as shown in Fig. 3 [19]. In the system, the 2-opt algorithm is driven by physical chaotic dynamics arisen from an analog circuitry.

The same benchmark problem as in section 3 was used. Figure 4 shows an example of the time evolution of the best solution where the system found the optimal solution at 378th iteration.

Table 1 compares the result from the exponential chaotic tabu search hardware and that from the Hopfield-type chaotic neuro-computer system with solution constructing method.

Figure 3: The exponential chaotic tabu search hardware system.

<table>
<thead>
<tr>
<th>Hardware System</th>
<th>Tai10a Benchmarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hopfield-Type</td>
<td>Optimal Solution</td>
</tr>
<tr>
<td>Chaotic Neuro-</td>
<td></td>
</tr>
<tr>
<td>Tabu Search</td>
<td></td>
</tr>
</tbody>
</table>
5. Chaotic Search Dynamics

Typical internal states of randomly chosen three neurons from the network when the above two systems successfully obtained the optimal solution are shown in Fig. 5 for the Hopfield-type chaotic neuro-computer with solution constructing technique, and in Fig. 6 for the exponential chaotic tabu search system, respectively. As shown in these figures, all of the neurons does not show fully developed chaos, but their dynamics are weak chaos, that is, the intermittent chaos near periodic window with period three or four. These observations suggest that global search and local search are chaotically interchanged for efficient chaotic search of the optimal solution without trapping in the local minima.

6. Conclusions

We have demonstrated the effectiveness of the physical chaotic neuro-dynamics to solve the quadratic assignment problems. We have used two hardware systems with analog chaotic neuron integrated circuits, which produce physical chaotic neuro-dynamics with continuous state variables. The first hardware system is the Hopfield-type chaotic neuro-computer system with solution constructing technique. In the second hardware system, the chaotic neuro-dynamics drives a tabu search algorithm realizing the exponential chaotic tabu search. We have solved the QAP as a benchmark with both systems, and compared the performance and the dynamics from these systems.

Acknowledgments

This study was supported in part by a Grant-in-Aid (No. 16300072) from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Figure 6: The time evolutions of the internal-state values, (a) $y_{1}(t)$, (b) $y_{2}(t)$, and (c) $y_{27}(t)$ of the exponential chaotic tabu search hardware.


Dynamic Behavior and Characteristics of the Modified ID model with Burst Firing

Shinya SUENAGA, Yoshihiro HAYAKAWA, and Koji NAKAJIMA

Laboratory for Brainware Systems, Laboratory for Nanoelectronics and Spintronics, Research Institute of Electrical Communication, Tohoku University Katahira 2-1-1, Aobaku Sendai 980-8577, Japan
Email: shinya@nakajima.riecl.tohoku.ac.jp, asaka@riecl.tohoku.ac.jp, hello@nakajima.riecl.tohoku.ac.jp

Abstract—We modify Inverse function Delayed (ID) model to implement burst firing. The ID model consists of two first-order differential equations and has an ability to oscillate such as a nerve-cell. Using the dynamics of oscillation, the ID model achieves the high performance in various applications of neural networks. The burst firing is other intrinsic dynamics of many nerve-cells and it has been assumed to have some effective information processing functions in a living neural network. If we implement burst firing property, which is more complex dynamics than the oscillation property of the ID model, to a neuron model, we may expect effective information processing. The modification of the ID model is based on the Hodgkin-Huxley model. The new model is able to be applied to solve the N-Queen problem, which is one of the optimization problems, and achieves higher performance. Furthermore, a prototype analog LSI of our model has been designed. We have confirmed the burst firing behaviours on the LSI by using simulation and measurement.

1. Introduction

Many neuron models, which have biological characteristics of the nerve cell at a certain level, have been proposed, e.g. the Inverse functionDelayed model [1] and the chaotic neural network model [2], because it is considered that newer functions arise from the neuron model to have biological characteristics.

As introducing biological dynamics, we focus on the burst firing over a neuron, because the relation between the functional role of the brain and this biological neural dynamic has been discussed so far. Burst firing is a phenomenon that occurs in many real neurons, and is characterized by repeated patterns of closely spaced action potentials. Such burst firing has been assumed to have some effective information processing functions [3, 4]. It is expected that the characteristics of burst firing add extra functions to artificial neural networks. A model which includes burst firing dynamics is proposed to extend the ID model. This extension is based on the relation among the ID model, BVP model, and Hodgkin-Huxley model. The proposed model has the ability to solve the combinatorial optimization problem. Hence, we can suggest a certain aspect of the effect of burst firing on information processing.

There are a few neurochips with burst firing ability, although the hardware implementation of a neuron model is widely investigated. Particularly, the neurochip with both the abilities to generate burst firing and to solve the optimization problems was not made so far. Additionally, LSI implementation is desired because an implemented chip makes it possible to perform fully parallel information processing.

This paper is organized as follows. In Section 2, we introduce the ID model. Then, the relation among the ID model, the BVP model, and Hodgkin-Huxley model is shown. Next, the burst ID model is proposed. In Section 3, we apply burst ID model network to the N-Queen problem. Finally, in Section 4, implementation and measurement of the neurochip of the burst ID model are shown.

2. Burst ID model

2.1. Inverse function Delayed (ID) Model

The ID model is represented by the following differential equations

\[ \tau_i \frac{du_i}{dt} = -u_i + Wx_i + \sum_{j \neq i} w_{ij} x_j - h_i \]  
\[ \tau_x \frac{dx_i}{dt} = u_i - g(x_i) \]

where \( x_i, u_i, h_i, n, W \) and \( w_{ij} \) are the output, the internal state, and the bias of the \( i \)-th unit, the total number of neural units, self-connection weight, and the synaptic weight from unit \( j \) to unit \( i \), respectively. \( \tau_u \) and \( \tau_x \) are time constants of internal state and output, respectively, and \( \tau_x < \tau_u \) is assumed. \( g \) is defined as the inverse output function. If \( \tau_x = 0 \) and \( g(x) = \tanh^{-1}(x) \), the ID model is equal to the Hopfield model.

The ID model’s Eqs. (1) and (2) can be reduced to the second-order differential equation:

\[ \tau_x \frac{d^2x_i}{dt^2} + \eta_i \frac{dx_i}{dt} = -\frac{\partial U_i}{\partial x_i} \]  

where

\[ \eta_i = \frac{\partial g(x_i)}{\partial x_i} + \frac{\tau_u}{\tau_x} \frac{\partial U_i}{\partial x_i} = \frac{1}{\tau_u} \left[ g(x_i) - Wx_i - \theta_i \right] \]

and where \( \theta_i = \sum_{j \neq i} w_{ij} x_j - h_i \) is the sum of inputs from other units and bias. Equation (3) represents a particle’s
motion on the potential $U$. In Eq. (3), the first term of the left-hand side corresponds to inertia, and then $\eta$ is a coefficient of the friction or resistance. $\eta$ is the nonlinear function of $x$.

If $g$ is an “N-shape” function, $\eta$ can be a negative value. For example, if the inverse output function $g$ is an “N-shape” function as follows:

$$g(x) = \frac{1}{\beta} \log \frac{x}{1 - x} + \alpha(x - \frac{1}{2}).$$

The coefficient $\eta$ has the region where it takes negative values, and then the ID model has an ability to oscillate.

If we put $\tau_g = c/b$, $\tau_x = 1/c$, $W = -1/b$, $\theta = a/b$, and $g(x) = x^2/3 - x$, the ID model is equal to the BVP model [5] without stimulus intensity. The BVP model is the reduced model of the well-known Hodgkin-Huxley (HH) model [6]. From this point of view, the ID model is one of many models which simulate nerve-cell activity.

2.2. Hodgkin-Huxley Model and the BVP Model

The HH equations describe the generation of an action potential. If the membrane potential $v$ is spatially constant and we omit the spatial derivative, the HH model for a squid giant axon is denoted as follows:

$$C \frac{dv}{dt} = G(v, m, n, h) + I_{\text{ext}}$$

$$\frac{dm}{dt} = \frac{1}{\tau_m(v)} (m^n(v) - m)$$

$$\frac{dn}{dt} = \frac{1}{\tau_n(v)} (n^n(v) - n)$$

$$\frac{dh}{dt} = \frac{1}{\tau_h(v)} (h^n(v) - h)$$

where $C$, $v$, $I_{\text{ext}}$, $G(v, m, n, h)$ are capacitance, membrane potential, constant current externally applied to the neuron, and membrane current, respectively. $m$, $h$, are the activation and inactivation variables of the Na$^+$ ionic channel, respectively. $n$ is the activation variable of the K$^+$ channel. These variables $m$, $n$, and $h$ take values between 0 and 1 and approach the sigmoid shape function $m^n(v)$, $n^n(v)$, and $h^n(v)$ with time constants $\tau_m(v)$, $\tau_n(v)$, and $\tau_h(v)$, respectively (see [6]).

The 4-dimensional HH equations can be reduced to 2-dimensional equations according to the two following approximations: $m = m^n(v)$ and $n = 0.8(1 - h)$ [7]. Applying these, we can eliminate the variables $m$ and $n$ and obtain the reduced equations:

$$C \frac{dv}{dt} = G(v, m^n(v), 0.8(1 - h), h) + I_{\text{ext}}$$

$$\frac{dh}{dt} = \frac{1}{\tau_h(v)} (h^n(v) - h).$$

On the $v-h$ phase plane, the nullcline of Eqs. (10) and (11) are a N-shape and a quasi-linear function, respectively. These topological characteristics are similar to the BVP model. Therefore, the ID model, which includes the BVP model, is the model in which neural cell dynamics are introduced.

2.3. Burst ID model

In this section, we construct a new bursting neuron model through the modification of the ID model. For this purpose, we consider the relation between the ID model and the HH model. The reduction described in the above section removes many dynamics, except for oscillation, from the HH model, and the variable $m$ is not equal to $m^n(v)$ in actual dynamics. Since $u$ and $x$ of the ID model correspond to $h$ and $v$ of the reduced HH equations, respectively, emergence of various dynamics that include burst firing can be expected by the introduction into the ID model of the another $z$ variable’s equation. The equation of this new variable $z$ is similar to the equation of $m$ and $z^n(x)$ is a sigmoid function such as $m^n(v)$. In this paper, we use $z^n(x) = \tanh(p(x - q))$. We add the linear term, $yz$, to the right hand side of Eq. (2). This term describe the effect of fluctuation depending on $z$.

Applying the above modification to the ID model, we propose the new model as follows:

$$\tau_x \frac{dx_i}{dt} = u_i - g(x_i) + yz_i$$

$$\tau_u \frac{du_i}{dt} = -u_i + W x_i$$

$$\tau_z \frac{dz_i}{dt} = -z_i + z^n(x_i) + \frac{1}{\gamma}$$

where $y, \theta$, and $\tau_z$ are a parameter and external input from other units, and the time constant, respectively. Time constants satisfy that $\tau_x < \tau_z < \tau_u$.

Although the right position of the external input $\theta$ is in Eq. (12) from the relation of the HH model, $\theta$ is given in Eq. (14) because of advantages over information processing, e.g. combinatorial optimization problems. The ID model also has the external input in Eq. (1) like a conventional network model, but in Eq. (2).

The extended ID model, which is described by Eqs. (12), (13), and (14), enters the burst firing state in a certain parameter region as shown in Fig. 1(a)-(c). This model has the burst firing ability which is shown in many nerve cells.

The mode of variable $x$ can be changed according to the value of external input $\theta$. If $\theta$ is small enough, the value of $x$ is stationary state at low value around $x = 0$. As $\theta$ increase, the state changes gradually to burst firing, oscillation, and stationary state at high value near $x = 1$. Hence, in the stationary state, the relation between input $\theta$ and output $x$ is similar to many neural network models, i.e. large and small input correspond to large and small output, respectively. Using this model, we can introduce the burst firing and oscillation dynamics into the transient between
values of burst ID model are low and experimentally.

where Hopfield model near the parameter set. The burst ID model shows the burst firing with this parameter set.

The burst ID model can denote much more performance than the ID model as shown in Fig. 2. Particularly, the difference of success rates at \( N = 6 \) is quite large. This result denotes the great potential of the burst ID model for information processing. To analyze the reason of this high performance is most outstanding issue for this model.

4. Hardware implementation

We implement the prototype neuro-chip of the proposed burst ID model. 1 Although the circuit is composed of a neuron module and a synapse module, we mainly describe the neuron module in this paper. We use analog technique to make these modules. This approach allow us to directly and faithfully implement the differential equations of the burst ID model. Figure 5 shows the microphotographs of whole neural network LSI.

4.1. Neuron module

To design a neuron module, the differential equations of our neuron are replaced by the electrical circuit. Equations (12), (13), and (14) are converted into following electrical equations:

\[
C_d \frac{dV_u}{dt} = I_u - I_u(x) + I_y
\]

\[
C_u \frac{dV_u}{dt} = -I_u + I_w x
\]

\[
C_y \frac{dV_y}{dt} = -I_{y_p} - I_{y^\infty(x)} + I_b
\]

where \( C_c \) and \( V_c \), \( * = x, u, y \), are capacitance and voltage of \( * \), respectively. \( y = -z \) is new variable for simplicity of circuit. In this circuit, \( y = -1 \) and \( W = -1 \). \( I_u \), \( I_u(x) \), and \( I_y \) are currents corresponding to \( u \), \( g(x) \), and \( y \) in Eq. (12), respectively. \( I_u \) and \( I_w \) are currents corresponding to \( x \) and \( W x \) in Eq. (13), respectively. \( I_{y_p} \) and \( I_{y^\infty} \) and \( I_b \) are currents corresponding to \( y^\infty \) and \( \theta \) in Eq. (13), respectively. All

1This work is supported by VLSI Design and Education Center(VDEC), the University of Tokyo in collaboration with Cadence Design Systems, Inc.
the terms of right hand side of Eqs. (17), (18), and (18) denote a current to the capacitance. These currents, except for $I_{g(x)}$ and $I_{g(x)}$, are easily realized by linear V-I converter using an OTA (Operational Transfer Amplifier).

$I_{g(x)}$ and $I_{g(x)}$, which are nonlinear functions of $V_x$, are comprised of several OTAs. $I_{g(x)}$ is a sigmoid function and comprise of a single OTA as shown in left bottom of Fig. 3. Circuit module for $I_{g(x)}$ is most complicated in neuron unit and is shown in the top of Fig. 3. This circuit is comprised of three OTAs which generate the left, the center, and the right parts of the “N-shape” waveform. The currents of these OTAs are combined using Kirchhoff’s law.

The circuit of whole neural module is figured as Fig. 3. We can obtain the good measurement result controlling adjusting voltages, although we observe some shift from the correct waveform. Figure 4 is the measurement result of this circuit which shows the burst firing. We can find a fluctuation in the interval between oscillation cycle because of noise and so on.

5. Conclusion

We have extended the ID model for the purpose of introducing the burst firing, which is one of the nerve-cell dynamics. Proposed neuron model, the burst ID model, indicates the burst firing with a certain external input region. It is expected that the additional new characteristics add extra functions to artificial neural networks, such as the burst firing plays some functional roles in the real brain. Actually, in the task of solving N-Queen problem, the burst ID model shows higher success rate than the original ID model does.

Additionally, the prototype chip for the burst ID model has been implemented and measured. We have obtained the result that our neurochip shows the burst firing waveform. It may be the first implementing trial of the neural network with the abilities of burst firing and processing the optimization problem. For the real time processing, the processing speed is important factor. It is necessary to make the special hardware for fully parallel processing. Because we can not fully use the capacity of neural networks by the simulation on ordinary computers.

References


An STDP-type Learning by Minimizing K-L Divergence for a Spiking Neural Network

Shigeo Sato, Kun Ma, and Koji Nakajima

Research Institute of Electrical Communication, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai 980-8577 JAPAN
Email: shigeo@riec.tohoku.ac.jp

Abstract—Spike timing dependent plasticity (STDP) has attracted much interest in research fields such as neuroscience or neurocomputing. We show that an STDP-type learning is obtained as the result of minimizing the Kullback-Leibler (K-L) divergence. A spiking neural network, where each neuron produces Poisson pulse series, has been confirmed to be able to learn by the proposed learning rule. We study the proposed learning rule together with successful simulation results.

1. Introduction

Experimental studies have observed synaptic potentiation when a presynaptic neuron fires shortly before a postsynaptic neuron, and synaptic depression when the presynaptic neuron fires shortly after. This dependence of synaptic modulation on the precise timing of two action potentials is known as spike timing dependent plasticity (STDP) [1]-[3]. STDP can be observed when the presynaptic and postsynaptic neurons fire within a 20-30 ms time window. These dynamical plasticity are believed to be important for learning in biological neural networks. Therefore, it is crucially informative to understand the basic principle of learning mechanism resulting STDP, both from the physiological and engineering viewpoints.

We study a learning rule for a spiking neural network. This study is motivated by two previously reported studies. First one is reported by Bohte and Mozer [4], in which STDP is obtained by minimizing entropy. Second one is reported by Xie and Seung[5], in which the reinforcement learning is obtained for a spiking neural network under the Poisson assumption related to the generation of the spike train. In the followings, first we introduce a spiking neuron model generating Poisson pulses. Second the learning rule is derived by minimizing the Kullback-Leibler (K-L) divergence. Finally, successful numerical results for learning is shown together with the confirmation of STDP.

2. Spiking Neuron

There have been proposed many spiking neuron models. Among them, we use the rather simple model proposed by Xie and Seung [5]. The total synaptic current received by neuron $i$ is given by the sum of the contributions over presynaptic neurons,

$$ I_i(t) = \sum_j w_{ij} h_{ij}(t), $$

where $w_{ij}$ is the synaptic weight from neuron $j$ to neuron $i$. The change in time of the synaptic current $w_{ij}h_{ij}(t)$ in neuron $i$ due to the spiking of neuron $j$ is given as,

$$ \tau \frac{dh_{ij}(t)}{dt} = \sum_a \delta(t - T^a_j) \zeta_{ij}^a - h_{ij}(t), $$

where $\tau$ is a time constant, and $T^a_j$ denotes the time of $a$th spike of neuron $j$. $\zeta_{ij}^a$ is a binary random variable modeling the stochastic release of neurotransmitters in response to a presynaptic spike. If the release event occurs, $\zeta_{ij}^a = 1$, otherwise $\zeta_{ij}^a = 0$. Please note that we use $\zeta_{ij}^a = 1$ always in our simulations. In order to adopt the Poisson assumption of neuron spiking, the relation between the synapse current $I(t)$ and the probability density of the neuron firing $f(I)$ is given as [4],

$$ f(I) = \frac{\beta}{\alpha} (\ln[1 + \exp(\alpha(\theta - I))] - \alpha(\theta - I)), $$

where $\theta$ is the firing threshold, and $\alpha$ and $\beta$ are free parameters. This relation has been derived for the stochastic variant of spike response model (sSRM) by incorporating the notion of a stochastic firing threshold [4],[6]. Figure 1 shows an example of $f(I)$ [5], where we use parameters such as $\alpha = 1/3, \beta = 60, \theta = 9.9$.

3. Learning Rule

As well known, the Kullback-Leibler divergence is a natural distance measure from a "true" probability distribution $R$ to an arbitrary probability distribution
$P$. The Hebb learning rule has been observed in biological neural networks, and is also derived by minimizing the K-L divergence for the Boltzmann machine[7], which is a stochastic neural network with symmetrical connections. Let us suppose that a network state is observed with a probability $P_\alpha$ and a target probability is given as $R_\alpha$, where $\alpha$ is the index of states. The K-L divergence is given as

$$G = \sum\limits_\alpha R_\alpha \ln \frac{R_\alpha}{P_\alpha}. \quad (4)$$

The updating rule for a synaptic weight is obtained as,

$$\Delta w_{ij} = -\eta \frac{\partial G}{\partial w_{ij}} = \eta \sum\limits_\alpha R_\alpha \frac{\partial P_\alpha}{P_\alpha} \frac{\partial P_\alpha}{\partial w_{ij}}. \quad (5)$$

where $\eta$ is a positive constant and should be chosen as enough small for successful gradient descent.

The network state is composed of all neuron states as $\sigma(t) = \{x_0(t), x_1(t), \ldots, x_n(t)\}$, where $x_i(t)$ is a binary variable and denotes the state of neuron $i$, i.e. $x_i(t) = 1$ if neuron $i$ fires at $t$. By considering the time course of the network state, the probability for the network to trace a pass $\Omega = \{\sigma(0), \sigma(\Delta t), \sigma(2\Delta t), \ldots, \sigma(T)\}$ is given as $P(\Omega) = \prod_{t=0}^{T} P(\sigma(t))$. Here, we assume that each neuron state is considered to be independent of other neuron states. This is rough assumption and not valid in general, but for mathematical convenience let us accept it here and verify the validness later. Now $P(\sigma(t))$ is given as the product of probabilities of all neuron states. By differentiating the natural logarithm of $P(\Omega)$, we obtain,

$$\frac{\partial \ln P(\Omega)}{\partial w_{ij}} = \sum\limits_{t=0}^{T} \ln P(x_i(t)). \quad (6)$$

Given the Poisson assumption on the generation of the spike train of each neuron, the probability for neuron $i$ to fire during the time interval $[t, t + \Delta t]$ is given as,

$$x_i(t) = \begin{cases} 1 & \text{with probability } p_i(t) = f_i(t) \Delta t, \\ 0 & \text{with probability } 1 - p_i(t), \end{cases} \quad (7)$$

which holds when $\Delta t$ is sufficiently small. By substituting $x_i(t)$ in Eq.(6), we obtain[5],

$$\frac{\partial \ln P(\Omega)}{w_{ij}} = \sum\limits_{t=0}^{T} \sigma_i(t) \frac{\partial \ln p_i(t)}{\partial w_{ij}} + (1 - \sigma_i(t)) \frac{\partial \ln(1 - p_i(t))}{\partial w_{ij}},$$

$$= \int_{0}^{T} f_i(t) \sigma_i(t) (s_i(t) - f_i(t)) h_{ij}(t) dt, \quad (8)$$

where $s_i(t) = \sum\limits_\alpha \delta(t - T_i^\alpha)$. Finally we obtain the updating rule of synaptic weights from Eq.(5),

$$\Delta w_{ij} = \eta \sum\limits_\alpha \frac{R_\alpha}{P_\alpha} \int_{0}^{T} f_i(t) \sigma_i(t) (s_i(t) - f_i(t)) h_{ij}(t) dt. \quad (9)$$

Figure 2: Poisson distribution of the firing rates.

4. Simulation Results

4.1. Learning

In order to check the learning rule derived in the above, a small network composed of 11 spiking neurons including a bias neuron is tested. A bias neuron triggers other neurons, and is crucial at the early stage of learning. This is because we set small values as initial synaptic weights, which are not enough for neurons to fire, then neurons tend to rest and no updating of synaptic weights occurs. Two target patterns (0000011111 and 1010101010), which have same probability ($R = 1/2$), are encoded to the average firing rates ($0 \rightarrow 100\text{Hz}, 1 \rightarrow 500\text{Hz}$). Please note that these rather high rates are used due to simulation convenience and can be adjusted for comparison with a
biological neuron by choosing proper parameters. For our case, \( f(I) \) is chosen as shown in Fig.1. After confirming that the Poisson pulses are generated based on Eqs.(1)-(3) as shown in Fig.2, the updating rule in Eq.(9) is applied to 10 neurons. Figure 3 shows the change of the K-L divergence as a function of learning epochs. It has been confirmed that the proposed learning works successfully. It can be seen also that certain non-zero error remains. This is caused by the stochastic property of pulses. The spike generation is stochastic and its interval has fluctuation related to the Poisson process. In the simulations, we allow \( \pm 20\% \) discrepancy of the average firing rate at decoding. However, as shown in Fig.3 the Poisson distribution is broad as it is characterized by the fact that its average and variance are equal, and thus the K-L divergence has certain non-zero value even if the learning completed.

\[
\Delta w_{ij} = \eta \sum \frac{R_\alpha}{P_\alpha} \left\{ \int_0^T \frac{f'_i(t)}{f_i(t)} s_i(t) h_{ij}(t) dt - \int_0^T \frac{f'_i(t)}{f_i(t)} f_i(t) h_{ij}(t) dt \right\},
\]

(10)

To simulate STDP, 3 neurons including a target neuron \( i \), a neuron \( j \) for input and a neuron \( k \) for biasing are required. The target neuron \( i \) has two synapses from neurons \( j \) and \( k \), and their amplitudes are set as \( w_{ik} > w_{ij} \). The bias neuron \( k \) and the synapse \( w_{ik} \) are both tuned for the target neuron to fire during the period \([0,100\,\text{ms}]\). By applying the proposed updating rule only to \( w_{ij} \), the spike timing dependency of the change of the synaptic weight \( w_{ij} \) is confirmed. If the neurons \( j \) and \( i \) fire at \( t_{\text{pre}} \) and \( t_{\text{post}} \), respectively, we calculate the time difference by \( t_{\text{pre}} - t_{\text{post}} \). Figure 4 shows the spike timing dependency on the spike timing. It can be seen that the spike timing dependent plasticity is simulated successfully.

\[ \text{Figure 3: Change of the K-L divergence in learning.} \]

\[ \text{Figure 4: Relation between the change of a synaptic weight and the spike timing.} \]

4.2. STDP

The dependence of synaptic modulation on the precise timing of two action potentials is known as spike timing dependent plasticity or STDP. STDP can be observed when the presynaptic and postsynaptic neurons fire within a 20-30 ms time window. Bohte and Mozer[4] have reported the STDP is realized by minimizing entropy with the sSRM. Our derivation in the previous section is done in the same manner. We, however, confirm by ourselves the synaptic modulation according to the derived learning rule. If a target neuron \( j \) fires at \( t_1 \), we obtain the following equation from Eq.(9).

\[ \Delta w_{ij} = \eta \sum \frac{R_\alpha}{P_\alpha} \left\{ \int_0^T \frac{f'_i(t)}{f_i(t)} s_i(t) h_{ij}(t) dt - \int_0^T \frac{f'_i(t)}{f_i(t)} f_i(t) h_{ij}(t) dt \right\}, \]

(10)

5. Conclusion

A learning rule for a spiking neural network has been derived by minimizing the Kullback-Leibler divergence. The numerical simulations show that the derived learning rule works well for a small neural network. Furthermore, the spike timing dependent plasticity has been confirmed also by the proposed learning. In the derivation of the learning rule, we use the
assumption that a neuron state is independent of other neurons. Though it is due to mathematical convenience and may be harmless for a small network like we have tested, such assumption is never true in general case. Studies in detail are required in future.

It is intriguing to compare the proposed learning with the Hebb learning. Both are obtained by minimizing the K-L divergence, the former for a spiking neural network while the latter for a conventional neural network. This analogy may be one of keys for understanding learning algorithms in a biological neural network.

Acknowledgments

This work was supported by a Grant-in-Aid from the Ministry of Education, Science, Sports, Culture and Technology, Japan.

References


Non-linear Model of a parallel hybrid power audio amplifier

Lorenzo Chiesi†, Emilio Lorenzani‡, Giovanni Franceschini† and Alberto Bellini‡

†DIE - University of Parma, Parco Area delle Scienze, 181/A, 43100, Parma, Italy
‡DISMI - University of Modena and Reggio Emilia, Via G.R. Fogliani, 1, 42100, Reggio Emilia, Italy
Email: giovanni.franceschini@unipr.it, bellini.alberto@unimore.it

Abstract—Parallel audio amplifier are hybrid amplifiers that couple a linear and a switching stage with a parallel connection. A proper control must be designed so that the voltage is controlled by the linear stage, while most of the current is driven by the switching stage. So doing the overall architecture achieves the typical efficiency of the switching stage and the distortion of the linear stage.

The control implementation is very critical because of the non-linear nature of the switching stage. Moreover the modulator of the switching stage is often a hystereris based self-oscillator, in order to achieve a simple and efficient architecture. The paper deals with the model of the hybrid audio amplifier architecture aiming at retriving the most critical items, and at defining simple relationships for the optimal design of the amplifier.

A prototype hybrid amplifier was realized, that confirms the effectiveness of the proposed method.

1. Introduction

An audio amplifier of small size and high efficiency can meet the needs of several markets, including consumer, medical and and military applications. The switching audio amplifier, commonly known as class D amplifier, allows achieving high efficiency and small size. The distortion level depends on the filter, the switching frequency, and the PWM technique employed, [1].

This paper presents a numerical analysis aimed at establishing the equations that can help in the design and lead to a better understanding of its behavior. Numerical analysis allows to aid in the proper selection of the components in the design of an hybrid audio amplifier with a parallel architecture. Numerical analysis provides the estimation of efficiency and switching frequency as a function of the input voltage, the system parameters, and stray parameters. Then efficiency and bandwidth can be predicted relying on simple relationship, retrieved from numerical analysis. A prototype was realized in order to verify retrieved relationships.

2. Modeling of parallel hybrid amplifiers

Combinations of analog and switching techniques have led to the development of audio power amplifiers capable of producing low distortion with high efficiency. To this aim at least two coupling connections are possible: parallel and series connections. In order to achieve a nice trade-off between efficiency and distortion, a class AB amplifier is chosen for the linear stage, and a MOSFET half-bridge is chosen for the switching stage. The choice of the class AB allows to tune distortion as needed. The parallel topology was suggested as the best option, because of its intrinsic capability of scaling the size of devices irrespectively of the delivered output power. Hence the investigation was limited to the parallel topology [3].

Due to the non-linear nature of the system simulations were used to define sizing relationship relying on an accurate non-linear model of the system. The major purpose is to characterize the behavior of the switching frequency, of delivered power and efficiency as a function of system parameters and of the delay of the control loop $T_d$ introduced by the stray parameters. Then benchmarks were made to seek for optimal values of system parameters in order to achieve the optimal values of efficiency, and power bandwidth. Moreover the effects of the delay of the control loop $T_d$ were analyzed for the same purposes. Simulations were made relying on MATLAB simulink tools, citeMatlab.

The parallel coupling of two voltage amplifier is a very critical issue. Each asymmetry between the stages that force the voltage on the same node may results in large currents, with a direct increase of power dissipation. It turns out that unavoidable voltage offsets, small gain variations, phase displacements and ripple may results in huge unexpected power dissipation.

The above issues can be tackled relying on an architecture where the former amplifier is a voltage amplifier that forces and controls the voltage on the load, while the latter is a current servoamplifier, controlled by the current delivered by the voltage amplifier, Fig. 1. Specifically the class AB stage is operated as a voltage amplifier, that forces and controls the voltage on the load, ensuring a very low distortion. The class D stage is operated as a current servoamplifier, whose current is $K_c$ times the current delivered by the class AB, $I_D = K_c I_{AB}$. Then current relationships are simply $I_{AB} = I_{spk} / (K_c + 1)$.

Provided that $K_c$ is large, most of the current is driven by the class D. Hence the efficiency is close to that of the switching stage. The current delivered by the class AB is a small fraction of the whole current and mitigates the intrinsic switching ripple.

The above configuration is achieved thanks to the scheme reported in Fig. 2.
The class D stage is composed by a half-bridge with MOSFET switches and an inductive filter. The current flowing on the inductance changes according to the operation of the MOSFET, that are driven by a Schmitt trigger that senses the current delivered by the class AB stage with a fixed threshold $I_T$. Specifically when the current delivered by the class AB overcomes $I_T$ the high-side MOSFET is switched on and the inductance is supplied by $+V_s$. Then the inductance current replaces the current delivered by the class AB on the load. The latter current is reduced until it falls below the threshold $-I_T$. At this time the low-side MOSFET is switched on and the inductance is supplied by $-V_s$. Then the inductance current is reduced. The excess current delivered by the switching stage flows in the linear stage. In summary the Schmitt trigger produces a self-oscillating mechanism that is kept even with no input signal, and a modulated signal whose switching frequency varies with the level of the modulating signal. The current delivered by the switching stage follows the current set-point with a ripple of amplitude $I_T$, that is compensated by the linear stage. The detailed structure embodies the so-called "current dumping" approach.

2.1. Self-oscillating switching behavior

Because of the self-oscillating structure the switching frequency is strongly varying with the level of the input signal. This is an attractive property from the point of view of an EMC-compliant design. Simulations show that the switching frequency varies with the level of the input signal, Fig. 3. With no input or low level signals the switching frequency is at its maximum level, then it decreases down to its minimum that is reached when the signal is close to the clipping level. The switching frequency is directly related to the rate of variation of the inductance current, that depends on the voltage applied and on the inductance value. In the former case the voltage applied on the inductance is the full voltage supply, while in the latter is reduced as much as the input signal is close to the supply level. Hence the maximum switching frequency $f_{SW,\max}$ can be easily modeled in the case of no input signal Eq. (1).

$$f_{SW,\max} = \frac{V_s}{4I_T L}$$

The maximum switching frequency is inversely proportional to the inductance value $L$. FFT analysis of sequential time windows were made with $T_w = 64 \mu s$ and a sampling frequency of $F_s = 2 \text{ MHz}$. Waterfall results for $I_D$ is reported in Fig. 4 assuming a sinusoidal input signal at $f = 1 \text{ kHz}$.

The maximum switching frequency is directly related to the power bandwidth, that is inversely proportional to the inductance value. Specifically the inductance limits the slew-rate of the switching stage. Then increasing the inductance the bandwidth at which the efficiency is at its maximum level decreases.

Simulations show that the effects of $I_T$ and $V_s$ on bandwidth are negligible, while the efficiency slightly decreases with $I_T$, as the class AB dissipation increases with $I_T$. On the other side while the efficiency increases with $V_s$, as the delivered power increases with the square of $V_s$ while the power dissipation increases linearly with $V_s$.

The above relationships define the so-called efficiency bandwidth, that is the bandwidth at which the efficiency is close to its maximum level. Ideally the efficiency bandwidth should include the frequency where statistically the audio signals is confined. It turns out that the efficiency bandwidth should be larger than a few kHz. Heuristically Eq. (2) is obtained from Fig. 5.

$$BW_{\eta,\max} \approx 0.1 \frac{Z_{Sp,k}}{L}$$  (2)
where $Z_{Spk}$ is the loudspeaker impedance, that is assumed constant with frequency for our purposes.

Fig. 3: Switching frequency versus the level of the input signal.

Moreover power dissipation and maximum efficiency can be predicted modeling the total power dissipation by the power dissipation of the class AB only, that always drives always a triangular wave current with amplitude $I_T$ even with no input signal.

MOSFET losses, iron losses, biasing current, stray parameters losses are not modeled here. The non-linear behaviour of the inductance is modeled from the magnetic characteristics of magnetic material used for coils inductance. Hence the maximum efficiency can be defined as Eq. (3) when the maximum power $P_{RMS}$ is delivered,

$$P_{RMS} = P_{Spk} = \frac{V_s}{2Z_{Spk}}$$

$$\eta_{max} = \frac{1}{Z_{Spk}I_T + V_s}$$

Eq. (4) is obtained replacing $I_T$ from Eq. (1) in Eq. (3).

$$\eta_{max} = \frac{1}{\frac{Z_{Spk}}{f_{SW Max}}L + 1}$$

Eq. (4) shows that the optimal trade-off between $f_{SW Max}$ and $L$ must be found in order to keep efficiency at its maximum level.

The above relationships allow to define a simple procedure for the sizing of the main parameters of the audio amplifier, in order to achieve optimal efficiency and distortion. Specifically a high switching frequency results in a simpler choice of system parameters in order to achieve a high efficiency in the audio bandwidth, i.e. in the bandwidth where most of the energy of the audio signals is confined.

Specifically in order to increase efficiency $I_T$ should be reduced, however it is constrained by Eq. (1). Then the only degree of freedom is the choice of $f_{SW Max}$. Provided that devices are chosen so that $f_{SW Max}$ can be very high, the inductance value $L$ is constrained by the bandwidth Eq. (2) and the maximum desired efficiency Eq. (4).

Eventually simulations show that the most critical parameter is the delay of the control loop $T_d$, that is mainly produced by intrinsic stray parameters of circuit implementation of current control and measurements. It turns out that the ideal behaviour, and attractive values of efficiency can be achieved, provided that a very fast implementation of the current sensing and of the Schmitt trigger is available. A control loop delay larger than 1μs prevents from a prompt feedback action and the desired switching correcting cannot be tracked. Reducing the inductance value the rate of variation of the current increases hence the same delay produces an higher current overshoot. The optimal design of circuit parameter $L$, $I_T$ can be made relying on a minimum value of $T_d$.

Figure 4: Waterfall of the switching current $I_D$.

Figure 5: $Z_{Spk} = 4\ \Omega$, $L = 100\ \mu H$, $I_T = 200\ mA$, $V_s = 40\ V$. Amplifier efficiency versus input signal frequency.

3. Experimental results

A prototype was realized in order to verify the relationships obtained by experiments. Its specification are

$$P_{RMS} = 150\ W, Z_{Spk} = 4\ \Omega, L = 150\mu H, V_s = 40\ V, I_T =$$
250 mA, $P_Q = 7$ W and full audio bandwidth operation [10 Hz, 22 kHz]. A commercial IC was used for the class AB stage, while discrete components were used for the class D stage. Measured current waveforms show that the current dumping behaviour is achieved in nice agreement with simulations, Fig. 7. These results are obtained thanks to a suitable current sensing design, that results in a delay of the control loop, $T_d \approx 500$ ns.

Figure 7: Experimental currents waveforms $Z_{S \text{pk}} = 4$ $\Omega$, $V_s = 40$ V, sinusoidal input with $f = 1$ kHz, $T_d < 1\mu$s, $I_{AB}$ (red line), $I_D$ (green line), $I_{SP \text{pk}}$ (blue line).

Fig. 8 report power absorbed, dissipated, and efficiency as a function of the power delivered to the load, with a sinusoidal input at 1 kHz and $Z_{S \text{pk}} = 4$ $\Omega$. Experiments are in nice agreement with simulation results. In fact the average power dissipation estimated with no input signal is of about 12 W, while the experiments report a value of about 14 W. Discrepancy between simulation and experiments is mainly caused by the lack of stray parameters, like MOSFET $R_{DS\text{on}}$ and inductance series resistance. Similarly the maximum efficiency computed by Eq. (3) is of about 97 %, while experiments report a value of about 85 %. The discrepancy can be partially justified by stray parameters. Signal conditioning circuitry, MOSFET drivers, MOSFET switching losses justify 7.5 W of power dissipation.

Fig. 9 reports the efficiency versus the input signal frequency, with $P_{S \text{pk}} = 70$ W and $Z_{S \text{pk}} = 4$ $\Omega$. The computed value of $BW_{\eta_{\text{max}}}$ is of 2.6 kHz Eq. (2) that is nicely close to the experimental result of about 3 kHz.

Figure 8: $f = 1$kHz, $Z_{S \text{pk}} = 4$ $\Omega$. Power and efficiency versus delivered power.

Figure 9: $P_{S \text{pk}} = 70$ W, $Z_{S \text{pk}} = 4$ $\Omega$. Efficiency versus input signal frequency.

4. Conclusion

Parallel hybrid configuration are promising architecture for power audio amplifiers, as they allow to achieve typical distortion figures of linear amplifiers together with typical efficiency of switching amplifiers.

The parallel hybrid configuration was modeled and simulated in order to define simple relationships useful to find critical parameters, and to aid design choices. A prototype amplifier was realized that confirms the effectiveness of the proposed approach.

References


Design-Oriented Hopf Bifurcation Boundary in Parallel-Connected Buck Converters Under Democratic Current-Sharing Control

Herbert H. C. Iu* and Chi K. Tse†

* School of Electrical, Electronic and Computer Engineering, The University of Western Australia, Australia
† Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong

Abstract—This paper studies a system of parallel-connected dc/dc buck converters under democratic current-sharing control. Under this control scheme, the reference current is generated by a typical proportional-integral (PI) controller for voltage regulation and the currents in individual converters are programmed by a built-in interleaving switching control. The effects of variations of practical parameters such as proportional gain, integral time constant, duty ratio and output power level are studied. Extensive simulations are performed to capture the Hopf bifurcation boundary which is the first instability boundary of practical importance. Unlike those in some previous studies which use abstract system parameters, the results in this paper are presented in terms of practical parameters and hence can be directly used for comparison of operating ranges for practical parallel-connected converters.

1. Introduction

Power supplies based on paralleling switching converters offer advantages over a single and high-power power supply. They enjoy low component stresses, increased reliability, increased power processing capability and improved fault tolerance [1]–[3]. Parallel converter systems become very popular in distributed power systems for front-end and load converters.

In parallel converter systems, mandatory control is needed to ensure proper current sharing, and many effective control schemes have been proposed [4]–[11]. One widely used method for balancing the currents is the democratic current sharing method [6], [8], [9]. In this method, a common reference current is generated from some form of feedback, and each converter’s inductor current needs to follow this reference current tightly. As a result, all modules share the load current equally.

In this paper, we attempt to study the stability of a system of parallel-connected buck converters under a democratic current-sharing control. We will study the first instability boundary, which is a Hopf bifurcation boundary, in terms of variations of some practical design parameters.

The paper is organized as follows. In Section 2, we present the control method of the parallel buck converter system and give the exact state equations that describe the system’s dynamics. In Section 3 we present the stability boundaries and some discussions about the results.

2. Operation of the Parallel Converter System

2.1. Current-Sharing Control Operation with Built-in Interleaving

Figure 1 shows the parallel system under study. It consists of two dc/dc converters which are connected in parallel feeding a common load. In this system, an appropriate current sharing scheme is needed to ensure that the current drawn by the load is shared properly between the two converters. In this paper, assuming that the parameters of the two converters are identical, and the current drawn by the load (output current) is shared equally. Also, a peak current-mode control scheme is employed. Denoting the two converters as Converter 1 and Converter 2, as shown in Fig. 1, the operation of the system can be described as follows. The converters are operating in continuous conduction mode. Both converters are controlled via a simple current-mode control scheme, in which the inductor current of Converter 1, i1, and that of Converter 2, i2, are compared with a reference current i_{ref} (which defines the peak inductor current) to generate the control signals that drive switches S_1 and S_2. Under our peak current-mode control scheme, after S_1 (S_2) is turned on, i_1 (i_2) ramps up until it reaches i_{ref}, then S_1 (S_2) is turned off and i_1 (i_2) ramps down. S_1 (S_2) remains off until it is turned on by the peri-
Figure 2: Steady-state inductor current waveforms under the built-in interleaving switching scheme, showing the switching sequence in each switching period \((d \leq 0.5)\).

odd clock signal. At the beginning of each switching period \(T\) (i.e., \(t = nT\)), only one switch is turned on. In our case, \(S_1\) and \(S_2\) are turned on alternately by the periodic clock signal, i.e., each switch is actually turned on every \(2T\). The scenario is illustrated graphically in Fig. 2.

The system is a closed-loop system. The reference current \(i_{ref}\) is generated from the output voltage through a typical proportional-integral (PI) controller. It is adjusted automatically such that the output voltage is regulated.

### 2.2. Exact State Equations

In this subsection, we derive the state equations for the system of parallel-connected buck converters. Using the state equations, simulation of cycle-by-cycle time-domain waveforms as well as bifurcation diagrams can be performed. Figure 1 shows the schematic of two buck converters connected in parallel. The presence of four switches \((S_1, S_2, D_1\) and \(D_2\)) allows a total of sixteen possible switch states, and in each switch state the circuit is a linear third-order circuit. The diodes \((D_1\) and \(D_2\)) are assumed to be ideal, with threshold equal to zero.

The system can be regarded as a variable structure that toggles its topology according to the states of the switches. When the converters are operating in continuous conduction mode, diode \(D_i\) is always in complementary state to switch \(S_i\), for \(i = 1, 2\). That is, when \(S_i\) is on, \(D_i\) is off, and vice versa. Hence, only four switch states are possible during a switching cycle, namely (i) \(S_1\) and \(S_2\) are on; (ii) \(S_1\) is on and \(S_2\) is off; (iii) \(S_1\) is off and \(S_2\) is on; (iv) \(S_1\) and \(S_2\) are off. The state equations corresponding to these switch states are generally given by

\[
\begin{align*}
\dot{x} &= A_1 x + B_1 E & \text{for } S_1\text{ and } S_2 \text{ on} \\
\dot{x} &= A_2 x + B_2 E & \text{for } S_1\text{ on and } S_2 \text{ off} \\
\dot{x} &= A_3 x + B_3 E & \text{for } S_1\text{ off and } S_2 \text{ on} \\
\dot{x} &= A_4 x + B_4 E & \text{for } S_1\text{ and } S_2 \text{ off},
\end{align*}
\]

\(A_i\) and \(B_i\) for the case of two buck converters are given by

\[
\begin{align*}
A_1 &= \begin{bmatrix} 1/C(R + r_c) & \frac{R}{L_i(R + r_c)} & \frac{R}{L_i(R + r_c) + r_{k1}} \end{bmatrix}, \\
B_1 &= \begin{bmatrix} \frac{1}{L_i} & 0 & 0 \end{bmatrix}, \\
A_2 &= \begin{bmatrix} 1/C(R + r_c) & \frac{R}{L_i(R + r_c)} & \frac{R}{L_i(R + r_c)} \end{bmatrix}, \\
B_2 &= \begin{bmatrix} \frac{1}{L_i} & 0 & 0 \end{bmatrix}, \\
A_3 &= \begin{bmatrix} 1/C(R + r_c) & \frac{R}{L_i(R + r_c) + r_{k1}} & \frac{R}{L_i(R + r_c) + r_{k1}} \end{bmatrix}, \\
B_3 &= \begin{bmatrix} \frac{1}{L_i} & 0 & 0 \end{bmatrix}, \\
A_4 &= \begin{bmatrix} 1/C(R + r_c) & \frac{R}{L_i(R + r_c) + r_{k1}} & \frac{R}{L_i(R + r_c) + r_{k1}} \end{bmatrix}, \\
B_4 &= \begin{bmatrix} \frac{1}{L_i} & 0 & 0 \end{bmatrix}.
\end{align*}
\]

where \(E\) is the input voltage, \(x\) is the state vector defined as

\[
x = [v_i \ i_1 \ i_2]^T,
\]

and the \(A\)'s and \(B\)'s for the case of two buck converters are given by

\[
B_1 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, \quad B_4 = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}.
\]

The reference current \(i_{ref}\) is derived from the feedback compensator. Here the compensator is a PI controller, i.e.,

\[
i_{ref}(s) = -K_p \left(1 + \frac{1}{Ts}\right)
\]

where \(i_{ref}(s)\) and \(V_c(s)\) are the Laplace transforms of \(i_{ref}(t)\) and \(v_c(t)\); \(v_c(t)\) is the error between the reference voltage \(V_{ref}\) and the output voltage \(v_o\); \(K_p\) and \(T_i\) are the parameters in the PI controller.

According to the feedback circuit (in Fig. 1), we can derive the state equation for \(i_{ref}\)

\[
\frac{di_{ref}}{dt} = -K_p \left(1 + \frac{1}{Ts}\right) v_o + \frac{K_i}{T_{F1}} v_{ref}.
\]

Also, the output voltage \(v_o\) can be written as

\[
v_o = v + r_{c} i_c = v + r_{c} (i_1 + i_2 - \frac{v_o}{R})
\]
1. Introduction

In this paper, a system of parallel buck converters under democratic current-sharing control has been studied. The system is a closed-loop system with the reference current generated from a typical PI controller. We study the effects of varying certain controller and system parameters on the system’s stability. Based on the simulations results, stability boundaries are constructed, which are useful for practical design.

2. System Description

The circuit under study is a parallel buck converter system with the reference current generated from a typical PI controller. We study the effects of varying certain controller and system parameters on the system’s stability. Based on the simulations results, stability boundaries are constructed, which are useful for practical design.

3. Simulation Results

In this section, simulation results of the parallel converter system are presented. As mentioned before, we perform computer simulations based on the exact state equations and the alternating switching rule on the operation of the system derived in Section 2. Moreover, we will study the effects of varying certain controller and system parameters on the stability of the system. The circuit parameters used in our simulations are shown in Table 1.

Depending on the various parameters, the system can exhibit either stable or unstable trajectories. In this case, due to the presence of the outer voltage loop which is bandlimited, the typical bifurcation with which the system loses stability is a Hopf-type bifurcation [11]. A typical trajectory of \( v \) vs \( i_1 + i_2 \) in the stable region (before Hopf bifurcation) is shown in Fig. 3. It is a stable \( T \)-periodic orbit and is the desired normal operation of practical power converters. A typical trajectory of \( v \) vs \( i_1 + i_2 \) in the unstable region (after Hopf bifurcation) is shown in Fig. 4. It is a quasi-periodic orbit.

First of all, we fix the output power \( P_o \) level at 14.4 W (\( v_o=12 \) V and \( R=10 \) Ω), and identify the stable region in the parameter space of \( K_1 \) and \( 1/T_{F1} \). Figure 5 shows the stability boundaries for different duty ratios. The duty ratio is given by \( v_o/E \) in our system. The area under the stability boundary corresponds to the stable region while the area above the boundary is the unstable region. The stable range of \( K_1 \) diminishes rapidly as \( 1/T_{F1} \) increases. We observe that for a smaller duty ratio, the stable region is larger.

Next, we fix the duty ratio at 0.3, and identify the stable region in the parameter space \( K_1 \) and \( 1/T_{F1} \). Figure 6 shows the stability boundaries for different output power levels. The output power is given by \( v_o^2/R \) in our system. For a higher output power level, the stable region is larger.

Then, we fix the duty ratio at 0.3, and identify the stable region in the parameter space \( K_1 \) and \( P_o \). Figure 7 shows the stability boundaries for different values of \( 1/T_{F1} \). For a smaller value of \( 1/T_{F1} \), the stable region is larger.

Finally, we fix \( 1/T_{F1} \) at 1000 s\(^{-1}\), and identify the stable region in the parameter space \( K_1 \) and \( P_o \). Figure 8 shows the stability boundaries for different values of the duty ratio. For a smaller duty ratio, the stable region is larger.

From the simulation results, we observe that increasing \( K_1 \), \( 1/T_{F1} \) and the duty ratio will decrease the size of the stable region. On the other hand, increasing the output power level will increase the size of the stable region.

4. Conclusion

In this paper, a system of parallel buck converters under democratic current-sharing control has been studied. The system is a closed-loop system, with the reference current generated from a typical PI controller. We study the effects of actual practical design parameters, such as circuit and controller parameters, on the system’s stability. Based on the simulations results, stability boundaries are constructed, which are useful for practical design.

Acknowledgment

This work was supported by supported by ARC Discovery Projects Grant (No. DP0559109), and Hong Kong Polytechnic University Research Grant (No. 1BB-ZA).
Figure 5: Hopf bifurcation boundary of $K_1$ vs $1/T_{F1}$ for different duty ratios. (Areas under the curves correspond to the stable regions.)

Figure 6: Hopf bifurcation boundary of $K_1$ vs $1/T_{F1}$ for different output power levels. (Areas under the curves correspond to the stable regions.)

Figure 7: Hopf bifurcation boundary of $K_1$ vs $P_o$ for different $1/T_{F1}$. (Areas under the curves correspond to the stable regions.)

Figure 8: Hopf bifurcation boundary of $K_1$ vs $P_o$ for different duty ratios. (Areas under the curves correspond to the stable regions.)

References


On the Application of a Server Type RTK-GPS to Electric Power Systems

Masashi Sakamoto† and Hitoshi Takata†
†Department of Electrical and Electronics Engineering, Kagoshima University
1-21-40 Korimoto, Kagoshima, 890-0065 Japan
Email: msakamot@Lime.ocn.ne.jp, takata@eee.kagoshima-u.ac.jp

Abstract—This paper is concerned with the application of a server type RTK-GPS to electric power systems. It enables us to observe the movements of electric power equipments accurately in real-time at a distant surveillant station. The experimental results indicate that this server type RTK-GPS can measure the three dimensional movements of poles with centimeter order’s accuracy.

1. Introduction

Electric power systems has repeatedly suffered from natural disasters such as typhoons, earthquakes, and tsunamis. Kagoshima in Japan is located in a typhoon path, thus its power systems have been destroyed by typhoons every year. In order to ensure the rapid restoration, we need to measure the accurate amount of change from normalcy for each installation.

In this paper we consider a surveying system to observe the movements of electric power equipments in real-time at a distant surveillant station by utilizing GPS (Global Positioning System)[1–8]. The GPS is the satellite navigation system operated by USA. Its cousins are called GLONASS by Russia, GALILEO by EU, and QZSS by Japan.

There are two types of GPS surveying methods, one of which is by code pseudo ranges and the other is by carrier phases. To acquire high-precision positioning, we make use of the method of carrier phase measuring. The carrier phase differential GPS is often used for precise positioning such as a geodetic surveys. Antennas at a reference station and of the other receivers take the signal data from the satellites simultaneously. Their data is processed to get the relative locations among observation points. By using information from the reference station, the original RTK-GPS (Real-Time Kinematic Global Positioning System) enables each roving receiver to determine its coordinates of the trajectory in real-time.

On the other hand, in this paper we explore a modification of RTK-GPS which has a server at a surveillant station. By this server, the coordinates of equipments’ movement is calculated by accepting the signal data from plural antennas, one of which is of the reference station and the other are of objective equipments. The signal data are transmitted by cables, networks, or wireless phones, etc. Hereafter, we simply call this system as a server type RTK-GPS.

The experimental results show that this server type RTK-GPS can measure the movements of poles with centimeter order’s accuracy at a surveillant station.

2. A Server Type RTK-GPS

We design a server type RTK-GPS whose conceptual scheme is shown in Fig.1.

Assume that $n$ satellites are caught by receivers $k$ and $m$ simultaneously. Let Satellite $p (p = 1)$ be a reference satellite and the others be Satellite $q \{q = 2, 3, \cdots , n\}$. The receiver $k$ of a reference station is set at known point on a standstill place, so its movement is supposed to be negligible small during the operation. The receiver $m \{m = 1, 2, \cdots , M\}$ are the objects whose movements are kept a careful watch on.

Each receiver directly transmits the raw data from the satellites to the server of surveillant station by cable, wireless phone, etc. Using these signal data, the server calculates the coordinates of objects’ movement with centimeter order’s accuracy.

Therefore, in this paper we make good use of the carrier phase to acquire high-precision positioning measurement[2,6,7].

The phase observable is the difference between the phase of the carrier signal transmitted by satellite and the phase of receiver oscillator. This difference in carrier phase can be regarded as the Doppler frequency shift that arises due to the relative motion between satellite and receiver. Then the Doppler frequency is integrated over the interval of the measurement period. Thus it is called the integrate carrier phase measurement, which is denoted by $\phi_k^p(t)$ for satellite $p$ and receiver $k$, where $t_c$ is the signal reception time measured by the receiver $k$.

Figure 1 also shows a conceptual view of the double difference, where two receivers $k$ and $m$ are located at the known point and the unknown point, respectively, so that the baseline vector between them is unknown. For the receivers located close together such that the separation is less than 20 km, ionospheric and tropospheric propagation delays are negligible small. Then the double difference phase observation becomes

$$
\Delta\phi_{km}^{pq}(t_c) = \phi_k^p(t_c) - \phi_m^p(t_c) - [\phi_k^n(t_c) - \phi_m^n(t_c)]
= \frac{f}{c} [\rho_k^p(t) - \rho_m^p(t)] - \frac{f}{c} [\rho_k^n(t) - \rho_m^n(t)] + \epsilon_{km}^{pq} + \epsilon_{km}^{pq}(t_c). \quad (1)
$$
Here $p_k(t)$ is the distance between the satellite $p$ and the receiver $k$. $t$ is the signal transmission time determined from GPS time. $N_{pq}^{km}$ is

$$N_{pq}^{km} = N_p^k + N_q^m - (N_p^k - N_q^m)$$

(2)

where $N_p^k$ is the unknown integer number associated with the ambiguity of carrier cycles at the initial time between the satellite $p$ and the receiver $k$. $f$ is the carrier frequency (1575.42MHz when L1), $c$ is the speed of light $(2.99792458 \times 10^8$ m/s), and $\varepsilon_{pq}^{km}$ is observation error.

The double difference data for the same set of $n$ satellites are available at both points.

We denote $p=1, q = 2, 3, 4, \cdots, n$ then the measurement can be formed:

$$\Delta(j) = [\Delta_{pq}^{km}(j), \Delta_{pq}^{km}(j), \cdots, \Delta_{pq}^{km}(j)]^T$$

(3)

where $t_r = j_0$, and $\delta_r$ is the data update interval of the receiver. In this positioning problem, we need to estimate three coordinates $(x_m, y_m, z_m)$ for the unknown position and $n-1$ values ($N_{km}^{12}, N_{km}^{13}, \cdots, N_{km}^{1n}$) corresponding to the integer ambiguity in (2). Therefore the state vector $\eta$ consists of $n + 2$ components as follows:

$$\eta(j) = [x^T(j), n^T(j)]^T$$

(4)

where

$$x(j) \triangleq [x_m(j), y_m(j), z_m(j)]^T$$

(5)

$$n(j) \triangleq [N_{km}^{12}(j), N_{km}^{13}(j), \cdots, N_{km}^{1n}(j)]^T$$

(6)

When all components are assumed to be constants, the state equation can be written as

$$\eta(j+1) = \eta(j) + \omega(j)$$

(7)

where $j$ is the discrete measurement time, and $\omega(j)$ is noise whose covariance is $Cov(\omega(j)) = Q(j)$.

The observation equation can be represented as follows:

$$\Delta(j) = h(\eta(j)) + \nu(j)$$

(8)

where $\nu(j)$ is measurement noise whose covariance is $Cov(\nu(j)) = R(j)$, and $h(\eta(j)) = [h_1(\eta(j)), \cdots, h_{q-1}(\eta(j)), \cdots, h_{n-1}(\eta(j))]^T$

(9)

$$h_{q-1}(\eta_j) = \frac{f}{c}[\sqrt{(x_k - x)^2 + (y_k - y)^2 + (z_k - z)^2} - \sqrt{(x_m - x)^2 + (y_m - y)^2 + (z_m - z)^2} - \sqrt{(x_m - x)^2 + (y_m - y)^2 + (z_m - z)^2} - \sqrt{(x_m - x)^2 + (y_m - y)^2 + (z_m - z)^2}] + N_{km}^{1q}$$

(10)

The next section will explain one of standard approaches of estimation for the above equations[6,7].

3. Estimation

3.1. Float Solution

The measurement equation is linearized by applying the Taylor expansion around $\hat{\eta}(j | j - 1)$ to Eq.(8) as follows.

$$\Delta(j) = h(\hat{\eta}(j | j - 1)) + C(j)(\eta(j) - \hat{\eta}(j | j - 1)) + \nu(j)$$

(11)

where $\hat{\eta}(j | j - 1)$ is prediction of $\eta(j)$ under $\Delta(0) \sim \Delta(j - 1)$

$$\hat{\eta}(j | j - 1) = [x^T(j | j - 1), \hat{\eta}^T(j | j - 1)]^T$$

(12)

$$C(j) = [H(j) \quad I_{n-1}], \quad H(j) \triangleq \left[ \frac{\partial h(\eta(j))}{\partial \eta(j)} \right]_{\eta(j)=\hat{\eta}(j | j-1)}.$$ 

For Eqs.(7) and (8), we have the extended Kalman filter:

$$\hat{\eta}(j | j) = \hat{\eta}(j | j - 1) + K(j)[\Delta(j) - h(\hat{\eta}(j | j - 1))]$$

(13)

$$\hat{\eta}(j + 1 | j) = \hat{\eta}(j | j), \quad \hat{\eta}(0 | -1) = \hat{\eta}(0)$$

(14)

$$K(j) = P(j)C^T(j)[C(j)P(j)C^T(j) + R(j)]^{-1}$$

(15)

$$P(j + 1) = P(j) - K(j)C(j)P(j) + Q(j)$$

(16)

where

$$\hat{\eta}(j | j) = [x^T(j | j), \hat{n}^T(j | j)]^T.$$

3.2. Fix Solution

We have obtained the estimates $\hat{x}(i \mid i)$ and $\hat{h}(i \mid i)$, in which the components of $\hat{h}(i \mid i)$ are real numbers, not integers. In order to perform the integer estimation, we accept the LAMBDA method at every epoch until they are fixed. The LAMBDA stands for Least squares AMBiguity Decorrelation Adjustment(3).

After the integers $n = \hat{n}(j | j)$ is fixed, $x(j)$ only is treated as the unknown vector at Eq.(4). Considering $n(j) = n$ being the known vector, the extended Kalman filter with respect to $x(j)$, which is a modification of Eqs.(13)- (16), is applied again to estimate the $\hat{x}(j | j)$.
Table 1: Reference station

<table>
<thead>
<tr>
<th>WGS84</th>
<th>WGS84</th>
</tr>
</thead>
<tbody>
<tr>
<td>(X,Y,Z)</td>
<td>(Lati.,Longi.,Height)</td>
</tr>
<tr>
<td>X = −3535582.769</td>
<td>B = 31°34′14″620N</td>
</tr>
<tr>
<td>Y = 4133321.417</td>
<td>L = 130°32′35″507E</td>
</tr>
<tr>
<td>Z = 3320002.240</td>
<td>H = 59.544</td>
</tr>
</tbody>
</table>

Table 2: Inclined distance (m)

<table>
<thead>
<tr>
<th>Direction</th>
<th>LDM</th>
<th>GPS</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>K1 → K2</td>
<td>0.66</td>
<td>0.67</td>
<td>0.01</td>
</tr>
<tr>
<td>K2 → K3</td>
<td>0.93</td>
<td>0.93</td>
<td>0.00</td>
</tr>
<tr>
<td>K3 → K4</td>
<td>2.44</td>
<td>2.52</td>
<td>0.08</td>
</tr>
<tr>
<td>K1 → K5</td>
<td>1.09</td>
<td>1.10</td>
<td>0.01</td>
</tr>
<tr>
<td>K5 → K6</td>
<td>0.75</td>
<td>0.73</td>
<td>0.02</td>
</tr>
<tr>
<td>K6 → K7</td>
<td>1.33</td>
<td>1.35</td>
<td>0.02</td>
</tr>
</tbody>
</table>

4. Experimental Results

We carry out the experiments on measuring the movement of poles by the server type RTK-GPS.

Both the serveillant and reference stations are set at EEEKU (Dep. of Electrical and Electronics Engineering of Kagoshima University) and they are connected by a cable. The receiver $k$ of reference station is put on the rooftop of EEEKU building whose coordinate point has already known (see Table 1). In the table, WGS84 is short for the World Geodetic System 1984. The server uses a personal computer (CPU: Pentium M 1.4 GHz) which includes a GPS analysis software RTKNav.

4.1. Experiment 1

The receiver $m$ as an object is located at a Taniyama training ground of KEPC(Kyushu Electric Power Co.) which is about 10km apart from EEEKU.

The GPS antenna of the receiver is tightly bounded on the top of a concrete pole which is easily inclined by a dump truck (see Fig.2).

The raw data received at this antenna is directly sent to the surveillant station by a transmitter, which is making good use of a wireless mobile phone (see Fig.3).

The movement of the pole is also observed a laser distance meter (LDM) in order to compare with our GPS surveying in accuracy.
A set of this experiment is designed and constructed by a collaboration with EEEKU and KEPC. The experimental results on March 15 in 2005 are disclosed below. Figure 4 shows the trace of \((x_m, y_m)\)-coordinate on a horizontal plane by an inclination of the pole of the receiver \(m\). K1 denotes a normal state of the pole, from which it leans to K2 → K3 → K4 and K5 → K6 → K7.

Table 2 tabulates the measured values of distance between two adjacent points \({K_i: i=1, \cdots, 7}\) by the LDM and the server type RTK-GPS, as well as their difference values.

Figures 5 and 7 are the time responses of the height \(z_m\) and the horizontal baseline \(r = \sqrt{(x_m - x_k)^2 + (y_m - y_k)^2}\), respectively. These results indicate that the server type RTK-GPS enables us to measure the movements of electric pole with centimeter order’s accuracy at the surveillant station being about 10km apart.

Next, we make the following test in order to check an experimental limit of accuracy of this method.

### 4.2. Experiment 2

The receiver \(m\) of an objective pole is set on the same rooftop of EEEKU building and its raw data is transmitted by a cable to the surveillant station. The distance between the receivers \(k\) and \(m\) is less than 10m. The experimental results on October 8 in 2005 are disclosed here as a typical one chosen from various tests. The pole is leaned from the normal state P1 to P2 → P3 → \cdots → P6. Figures 7 and 8 are the results of height \(z_m\) and baseline \(r\), and they are quite similar to Figs.5 and 6, respectively. Since the accuracy of this experiment is about 1.5cm, the limit of this method seems to be centimeter order.

### 5. Conclusions

We have considered a server type RTK-GPS, which is a modification of RTK-GPS having a server at a surveillant station. This system might be applied to observe the movement of electric power equipments in real-time under a distant surveillant station. The experimental results have been shown that we could measure the inclination of poles with centimeter order’s accuracy at the station. Future research is required to clarify the influence of noise, error, distance, receiver, and so on.

### Acknowledgments

The authors would like to give their sincere thanks to Kyushu Electric power Company for various supports in this research. These authors also express the gratitude to T. Hachino, S. Kawaji and T. Yamasaki of Kagoshima University, and GNSS Technologies Inc. for their helps to complete these experiments.

### References


Effects of Characteristics of Band-pass Filter to Operation of Class DE Inverter

Motoki Katayama†, Hiroyuki Hase†, Hiroo Sekiya†, Janming Lu†, and Takashi Yahagi†

†Graduate School of Science and Technology, Chiba University
1-33 Yayoi, Inage, Chiba, Chiba, 263-8522 JAPAN
Email: motoki.katayama@graduate.chiba-u.jp

Abstract—This paper investigates the effects of characteristics of band-pass filter to operation of class DE inverter. Two kinds of band-pass filter are applied to the design of class DE inverter, and circuit experiments are carried out. As the result, output voltage characteristic of the class DE amplifier with band-pass filter is very flat. Because of output characteristics of band-pass filters are flat at the nominal frequency. Moreover, when the operating frequency decreases from nominal frequency, both the class DE inverter with band-pass filters achieve ZVS.

1. Introduction

Class DE inverter[1]–[6] achieves high power conversion efficiency under a high frequency operation because of class E switching [7]. Class E switching means that both the switching voltage and the slope of switching voltage are zero when each switch turns on.

Class DE inverter presented in the previous papers includes a resonant circuit to acquire sinusoidal output. The output voltage is quite sensitive to the variations of operating frequency. Moreover, class E switching conditions are strict, so not only the class E switching conditions but also zero voltage switching (ZVS) cannot be achieved against the variations of operating frequency [6]. The high sensitivities of output power and switching conditions to the operating frequency are problems for class DE inverter.

Recently, class DE amplifier with band-pass filter is proposed in [8]. By using band-pass filter, it is confirmed the following two advantages. One is that the sensitivity of output voltage to the operating frequency can be suppressed. The other is that the ZVS can be satisfied when the operating frequency decreases from the nominal operation. However, it has never discussed that the effects of output characteristics of the filter to characteristics of the output voltage.

This paper investigates the effects of characteristics of band-pass filter to operation of class DE inverter. Two kinds of band-pass filter are applied to the design of class DE inverter, and circuit experiments are carried out. As the result, output voltage characteristic of the class DE inverter with band-pass filter is very flat. It is because that output characteristics of band-pass filters are flat at the nominal frequency. Moreover, when the operating frequency decreases from nominal frequency, both the class DE inverter with band-pass filters achieve ZVS.
2. Circuit Description

2.1. Class DE inverter with band-pass filter[8]

Figure. 1 (a), (b) shows the circuit topology of class DE inverter with resonant filter, and the one with band-pass filter[8], respectively. This paper presents class DE inverter with band-pass filter as shown in Fig. 1 (b). In this inverter, band-pass filter \( L_1 - C_1 - L_2 - C_2 \) is used as shown in Fig. 1 (b) instead of the resonant filter \( L_0 - C_0 \) in Fig. 1 (a). The band-pass filter passes frequencies with a certain bandwidth. If we set the bandwidth to cut the higher harmonics and direct component, pure sinusoidal output voltage is derived. It is recognized that by using the band-pass filter, the sensitivity of output voltage to the operating frequency can be suppressed, and class DE inverter with band-pass filter satisfy ZVS when operating frequency decreases. Moreover, since the circuit is satisfied with class E switching conditions, it achieves high power conversion efficiency under high frequency operation. However, it has never discussed that the effects of output characteristics of the filter to characteristics of the output voltage.

2.2. Investigation of the effects of characteristics of band pass filter.

In this paper, to find out the relation between output characteristics of the band-pass filter and characteristics of the output voltages, two band-pass filters which have different output characteristics are applied to class DE inverter. Figure 3 shows each output characteristics of two band-pass filters and resonant filter. Figure 3 (b) shows each output characteristics when the range of operating frequency is 0.95MHz to 1.05MHz.

It is expected that the output characteristics against the operating frequency depends on the flatness of the output characteristics of the output filter, which has never been confirmed yet. If it is true, it is possible to design the class DE inverter with various output characteristics by changing output filter. We recognize it is an important and useful guideline for designers.

3. Circuit Design

3.1. Assumptions

In order to derive the waveforms of the inverter, we give the following assumptions.

i. The switching devices, namely MOSFETs including anti-parallel diode have zero switching times, large off resistance enough to neglect the current through the switches, and nonzero on resistance. On resistances of the MOSFETs and the anti-parallel diodes are stated as \( r_\text{s} \) and \( r_\text{ds} \), respectively. In this paper, we use a same kind of the MOSFETs for the switches. Therefore, it is assumed that a value of on resistance of one MOSFET is identical that of another MOSFET.

ii. Shunt capacitance of each switching device, namely \( C_{\text{s}1} \) and \( C_{\text{s}2} \) include switch device capacitance. The equivalent series resistance of each capacitances can be neglected.

iii. In this paper, we use the second order constant K filter as a band-pass filter.

iv. All passive elements including switch on resistors operate as linear elements.

3.2. Parameters

We define the following parameters of the circuit in order to express circuit equations.

1. \( \omega = 2\pi f \) : The operating angular frequency.
2. \( X = \omega C_S = \omega (C_{S1} + C_{S2}) \)
3. \( \alpha = f_w / f_c \) : The ratio of bandwidth to center frequency of band-pass filter.
4. \( \beta = f / f_c \) : The ratio of operating frequency and center frequency of band-pass filter.
5. \( Z \) : The impedance of band-pass filter when band-pass filter is second order constant K filter.
6. \( D = D_{S1} = D_{S2} \) : The switch on duty ratio.
It is inevitable that two switch is same inverter are on state simultaneously for $D > 0.5$. This causes short circuit and breaks the inverter. Moreover, the dead time must be needed in order to achieve Class E switching. Therefore, the range of $D$ is determined as $0 < D < 0.5$.

### 3.3. Circuit Equations

Following above assumptions, the equivalent circuit of the proposed inverter is shown in Fig. 3. We consider operations at $0 \leq \theta < 2\pi$ to design circuit, where $\theta = \omega t$ presents the angular time. Using the parameters, the circuit equations are expressed as follows:

$$
\begin{align*}
\frac{dv_{s1}}{d\theta} &= -\frac{1}{X} \left( i_2 + \left( \frac{1}{R_{s1}} + \frac{1}{R_{s2}} \right) v_{s1} - \frac{V_D}{R_s} \right) \\
\frac{dv_1}{d\theta} &= \frac{\beta}{Z} \left( i_2 - i_1 - v_1 \right) \\
\frac{dv_2}{d\theta} &= \frac{\beta}{Z} \left( i_2 - i_1 - v_1 \right) \\
\frac{dv_1}{d\theta} &= \frac{\alpha}{Z} \left( i_2 - i_1 - v_1 \right) \\
\frac{dv_2}{d\theta} &= \frac{\alpha}{Z} \left( i_2 - i_1 - v_1 \right)
\end{align*}
$$

(1)

In (1), $R_{s1}$ and $R_{s2}$ are the equivalent resistance of MOSFETs $S_1$ and $S_2$, respectively. When the switch voltage is to be negative, the anti-parallel diode including the MOSFET turns on.

For the decision of design parameters, some techniques for numerical calculation are applied. Specifically, applying Runge-Kutta method and Newton’s method, the unknown parameters can be found, and the design values, that is, $X$ and $\beta$ are determined. The procedure for calculations of Newton’s method is the same as in [9].

### 4. Experimental Results

In this section, circuit experiments are shown. This paper apply two band-pass filter which have different characteristic. One of the design parameters of the class DE inverter with band-pass filter are $\alpha = 0.7$, the impedance of band-pass filter $Z = 5.0$, output resistance $R = 40$, the other ones are $\alpha = 0.9, Z = 3.5, R = 60$. The output characteristics of both filters shown in Fig. 3. The element values of each experiments are determined as shown in Tab. 1 and 2.

Figure 4 shows the examples of experimental waveforms by changing the operating frequency. Figure 4 (a-1), (b-1) shows the experimental waveforms for $f = 1.0MHz$. From these figures, we can confirm the operation of $v_{s1}$ achieves class E switching in both filters. In this paper, we define the state in Fig. 4 (a-1), (b-1) as nominal state. In the nominal state, each inverter achieves 90.4 % and 88.6% power conversion efficiency in Fig. 4 (a-1), (b-1), respectively. Figure 4 (a-2), (b-2) shows the waveforms for $f = 0.95MHz$, which is smaller than the nominal operating frequency. From these figures, the waveform of the switch $S_1$ never achieves class E switching conditions, but is satisfied with ZVS in both filters.

Table 1: Design values when design parameters are $\alpha = 0.7, Z = 5.0, R = 40$.

<table>
<thead>
<tr>
<th></th>
<th>Calculated</th>
<th>Measured</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>0.496$\mu$H</td>
<td>0.496$\mu$H</td>
<td>0.00%</td>
</tr>
<tr>
<td>$L_2$</td>
<td>1.013$\mu$H</td>
<td>1.012$\mu$H</td>
<td>0.10%</td>
</tr>
<tr>
<td>$C_{s1}$</td>
<td>4.52nF</td>
<td>4.50nF</td>
<td>-0.44%</td>
</tr>
<tr>
<td>$C_{s2}$</td>
<td>4.52nF</td>
<td>4.51nF</td>
<td>-0.22%</td>
</tr>
<tr>
<td>$C_1$</td>
<td>40.51nF</td>
<td>40.49nF</td>
<td>-0.05%</td>
</tr>
<tr>
<td>$C_2$</td>
<td>19.85nF</td>
<td>19.84nF</td>
<td>-0.05%</td>
</tr>
<tr>
<td>$R$</td>
<td>40.0$\Omega$</td>
<td>39.9$\Omega$</td>
<td>-0.03%</td>
</tr>
<tr>
<td>$D$</td>
<td>0.25</td>
<td>0.25</td>
<td>0.00%</td>
</tr>
<tr>
<td>$f$</td>
<td>1.0MHz</td>
<td>1.0MHz</td>
<td>0.00%</td>
</tr>
<tr>
<td>$V_c$</td>
<td>20.0V</td>
<td>20.0V</td>
<td>0.00%</td>
</tr>
<tr>
<td>$r_s$</td>
<td>0.16$\Omega$</td>
<td>0.16$\Omega$</td>
<td>0.00%</td>
</tr>
<tr>
<td>$r_D$</td>
<td>1.54$\Omega$</td>
<td>1.54$\Omega$</td>
<td>0.00%</td>
</tr>
<tr>
<td>$V_o$</td>
<td>9.37V</td>
<td>9.24V</td>
<td>-1.39%</td>
</tr>
<tr>
<td>$\eta$</td>
<td>92.7%</td>
<td>90.4%</td>
<td>-2.48%</td>
</tr>
</tbody>
</table>

Table 2: Design values when design parameters are $\alpha = 0.9, Z = 3.5, R = 60$.

<table>
<thead>
<tr>
<th></th>
<th>Calculated</th>
<th>Measured</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_1$</td>
<td>0.45$\mu$H</td>
<td>0.45$\mu$H</td>
<td>-0.89%</td>
</tr>
<tr>
<td>$L_2$</td>
<td>0.557$\mu$H</td>
<td>0.56$\mu$H</td>
<td>0.54%</td>
</tr>
<tr>
<td>$C_{s1}$</td>
<td>4.89nF</td>
<td>4.88nF</td>
<td>-0.20%</td>
</tr>
<tr>
<td>$C_{s2}$</td>
<td>4.89nF</td>
<td>4.91nF</td>
<td>0.41%</td>
</tr>
<tr>
<td>$C_1$</td>
<td>45.47nF</td>
<td>45.45nF</td>
<td>-0.04%</td>
</tr>
<tr>
<td>$C_2$</td>
<td>36.83nF</td>
<td>36.82nF</td>
<td>-0.03%</td>
</tr>
<tr>
<td>$R$</td>
<td>60.0$\Omega$</td>
<td>60.0$\Omega$</td>
<td>0.00%</td>
</tr>
<tr>
<td>$D$</td>
<td>0.25</td>
<td>0.25</td>
<td>0.00%</td>
</tr>
<tr>
<td>$f$</td>
<td>1.0MHz</td>
<td>1.0MHz</td>
<td>0.00%</td>
</tr>
<tr>
<td>$V_c$</td>
<td>20.0V</td>
<td>20.0V</td>
<td>0.00%</td>
</tr>
<tr>
<td>$r_s$</td>
<td>0.16$\Omega$</td>
<td>0.16$\Omega$</td>
<td>0.00%</td>
</tr>
<tr>
<td>$r_D$</td>
<td>1.54$\Omega$</td>
<td>1.54$\Omega$</td>
<td>0.00%</td>
</tr>
<tr>
<td>$V_o$</td>
<td>8.89V</td>
<td>8.73V</td>
<td>-1.80%</td>
</tr>
<tr>
<td>$\eta$</td>
<td>88.6%</td>
<td>88.2%</td>
<td>-0.45%</td>
</tr>
</tbody>
</table>
526

Figure 4: Numerical waveforms and experimental waveforms. (a) Design values when design parameters are $\alpha = 0.7, Z = 5.0, R = 40$. (b) Design values when design parameters are $\alpha = 0.9, Z = 3.5, R = 60$. (1) Operating frequency 1.00MHz, (2) Operating frequency 0.95MHz. Vertical: $D_v$: 5V/div, $v_S$: 20V/div, $v_o$: 20V/div. Horizontal: 400ns/div

by using band-pass filter instead of resonant filter, and it is similar the output characteristic of band-pass filter to the output voltage. This means that using flatter characteristic filter to class DE inverter, the sensitivity of output voltage to the operating frequency can be more suppressed.

5. Conclusion

This paper has investigated the effects of characteristics of band-pass filter to operation of class DE inverter. Two kinds of band-pass filter are applied to the design of class DE inverter, and circuit experiments are carried out. As the result, output voltage characteristic of the class DE amplifier with band-pass filter is very flat. Because of output characteristics of band-pass filters are flat at the nominal frequency. Moreover, when the operating frequency decreases from nominal frequency, both the class DE inverter with band-pass filters achieve ZVS.

References


A Novel Control Method of a Regenerative Power Storage System for Energy Saving of an Electric Machinery

Kaoru Inoue, Kenji Ogata and Toshiji Kato

Department of Electrical Engineering, Doshisha University, Kyotanabe, Kyoto 610-0321, JAPAN
Email: inoue@kairo.doshisha.ac.jp

Abstract— This paper proposes a novel control method of a regenerative power storage system for electric machinery to save the energy. The regenerative power is stored and then provided to a motor by the regenerative power storage system that consists of an electrolytic capacitor and a bi-directional DC-DC converter. A new mathematical model based on a state-averaging approach is derived. The obtained mathematical model is a bi-linear system, however it can be treated as a linear system with a control input and an external disturbance because one of the variables can be measured accurately by using a voltage sensor. The effectiveness of the proposed method is illustrated by both a simulation and an experiment.

1. Introduction

When a motor speed is reduced by using an electromagnetic brake, the mechanical energy of rotation is converted to the electrical power [1, 2, 3, 4, 5, 6]. This is called regenerative power. Because the saving of energy is advanced due to the increment of electric power demand and environmental problems, some efficient methods have been reported. For example, the regenerative power is returned to the power supply system, and the regenerative power is stored and provided by the regenerative power storage system composed by a bi-directional DC-DC converter and an electrolytic capacitor, chemical battery or flywheel system, etcetera. The former method is not applicable in Japan because of the regulation and the circulating current appears [10]. The latter method is suitable for electric machinery. In order to control the DC-DC converter, PID based control method is widely adopted based on transfer function models [2, 3, 4, 5, 6]. However, theoretical stability analysis of the closed-loop system is not well discussed.

This paper treats the regenerative power storage system composed by an electrolytic capacitor and a bi-directional DC-DC converter. A new mathematical model based on a state-averaging approach is derived. The obtained mathematical model is a bi-linear system, however it can be treated as a linear system with a control input and an external disturbance because one of the variables can be measured accurately by using a voltage sensor. We propose a novel control method of the regenerative power storage system to achieve an efficient motor drive system. Proposed control method has both a feedback and a feed-forward path. The effectiveness of the proposed method is illustrated by both a simulation and an experiment.

2. Objective system

2.1. System configuration

Figure 1 illustrates the objective system. A regenerative power storage system and a three phase inverter are connected by a DC link. A vector control method [7, 8] is employed in order to control a torque instantaneously in the case of acceleration and regeneration of the induction motor IM. The regenerative power is stored by the regenerative power storage system that consists of an electrolytic capacitor $C_1$ and a bi-directional DC-DC converter. The bi-directional DC-DC converter is composed of an inductor $L$ and IGBT switches SW1 and SW2. These IGBTs are complementary switches. When the regenerative power is stored by $C_1$, the DC-DC converter boosts the voltage. On the other hand, it works as a buck converter and provides the stored power to the motor when the motor is accelerated.

2.2. State-averaged model

Figure 2 shows the regenerative power storage system and the DC link capacitor $C_2$. This circuit has
two circuit modes illustrated by Fig. 3. It is assumed that IGBT switch and the freewheel diode have the same ON resistance \( r(\Omega) \) approximately.

Let us derive a model by means of the state space averaging approach [9]. The circuit mode a1 is modeled by

\[
L \frac{d}{dt} i_L = -ri_L + V_{DC2} - V_{DC1} \quad (1) \\
C_2 \frac{d}{dt} V_{DC2} = i_{DC} - i_L. \quad (2)
\]

Similarly, the circuit mode a2 is shown by the following equations.

\[
L \frac{d}{dt} i_L = -ri_L + V_{DC2} \quad (3) \\
C_2 \frac{d}{dt} V_{DC2} = i_{DC} - i_L. \quad (4)
\]

SW1 conducts a fraction \( D \) of each switching cycle and SW2 conducts for the remainder \( 1 - D \). A state variable is defined by

\[
x = \begin{bmatrix} i_L \\ V_{DC2} \end{bmatrix}. \quad (5)
\]

Then we obtain a new state-averaged model

\[
\frac{d}{dt} x = Ax + bDV_{DC1} + d 
\]

\[
A = \begin{bmatrix} -\frac{1}{L} & \frac{1}{C_2} \\ -\frac{1}{C_2} & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ -\frac{1}{C_2} \end{bmatrix}, \quad d = \begin{bmatrix} \frac{V_{DC1}}{C_2} \end{bmatrix}. \quad (7)
\]

The state-averaged model (6) is a bi-linear system because of the variables \( D \) and \( V_{DC1} \). However it can be seen that the term \( bDV_{DC1} \) is a function of \( D \) because \( V_{DC1} \) varies slowly comparing to \( D \) and it can be measured accurately by using a voltage sensor. Therefore, Eq. (6) is treated as a linear time-invariant system with a control input \( D \) and an external disturbance \( d \). In addition, \( d \) also can be measured accurately by means of a current sensor.

3. Proposed control method

3.1. Outline of the method

When the motor speed is reduced, the DC link voltage \( V_{DC2} \) increases because of the regenerative power. The voltage variation of \( V_{DC2} \) is undesirable, hence the regenerative power is stored by \( C_1 \) through the bi-directional DC-DC converter in order to keep \( V_{DC2} \) constant. This control problem is achieved by the following method; the DC link voltage \( V_{DC2} \) and an inductor current \( i_L \) are controlled to approach a constant reference voltage \( V_{DC2}^* \) and DC link current \( i_{DC} \), respectively.

This method is applicable in the case of the motor acceleration. The stored power in \( C_1 \) is provided with the motor by means of the control method. Consequently an efficient drive system will be achieved.

3.2. Control input

We define an ideal reference state of the system as

\[
x^* = \begin{bmatrix} i_{DC} \\ V_{DC2}^* \end{bmatrix} \quad (8)
\]

where the DC link current \( i_{DC} \) is time-varying and can be measured by using a current sensor accurately. The reference DC link voltage \( V_{DC2}^* \) is given appropriate constant voltage. An ideal reference system is given by the equation.

\[
\frac{d}{dt} x^* = Ax^* + bDV_{DC1}^* + d \quad (9)
\]

Let us subtract Eq. (9) from Eq. (6), then an error system is obtained as follows.

\[
\frac{d}{dt} x_e = Ax_e + bDV_{DC1} - bDV_{DC1}^* \quad (10)
\]

\[
x_e = x - x^* \quad (11)
\]

When a control input \( D \) is chosen as

\[
D = \frac{1}{V_{DC1}} (-K_ax_e + D^*V_{DC1}^*) \quad (12)
\]

\[
K_a = [K_{ai} \ K_{av}] \quad (13)
\]

and is applied to Eq. (10), a closed-loop system is obtained.

\[
\frac{d}{dt} x_e = (A - bKa)x_e \quad (14)
\]
4. Simulation and experiment

4.1. Setup

In order to illustrate the effectiveness of the proposed control method, a simulation and an experiment are performed. In the simulation, MATLAB simulink and simpowersystems are used. Table 1 shows the parameters setup. Gains $K_{ai}$ and $K_{av}$ are set to satisfy $0 \leq D \leq 1$. The reference DC link voltage $V_{DC1}^*$ is set 320 (V). Figure 4 shows the experimental system setup. The DC link voltage $i_{DC}$ is required to calculate the control input (12). Switching noise and its harmonics appears in $i_{DC}$. Hence, it is expected that more energy will be regenerated when the torque is decreased appropriately. This is one of the future work.

4.2. Simulated result

After accelerating the induction motor up to 1600 (rpm), the motor speed is reduced (regenerative brake) and accelerated by constant torque induced by the vector control. Figure 5 illustrates a simulation result where Speed and $T_e$ denoting the rotating speed of the motor and the torque, respectively. The motor is accelerated in the case of positive $T_e$, the motor works as the generator in the case of negative $T_e$. The inductor current $i_L$ and the DC link current $i_{DC}$ increase under the regenerative brake.

After 0.3(sec), the regenerative brake starts by means of the constant torque $T_e = -10$ (Nm). The regenerative power is stored by the capacitor $C_1$ through the bi-directional DC-DC converter, consequently the capacitor voltage $V_{DC1}$ rises. The DC link voltage $V_{DC2}$ coincides with the reference voltage $V_{DC2}^* = 320$ (V), $i_L$ is controlled to the reference $i_L^* = i_{DC}$.

4.3. Experimental result

Figure 6 shows an experimental result. After 0.3(sec), the Speed begin to decrease due to the regenerative brake. The capacitor voltage $V_{DC1}$ rises up around 0.38(sec), however, $V_{DC2}$ begin to decrease through the peak voltage 342 (V), although the torque is still kept constant $T_e = -10$ (Nm). A cause of the phenomenon is that the power consumption inducing $T_e = -10$ (Nm) exceeds the regenerative power. Hence, it is expected that more energy will be regenerated when the torque is decreased appropriately. This is one of the future work.
to around 340(V) because the regenerative power is stored by the capacitor $C_1$. The DC link voltage $V_{DC2}$ is controlled to the reference voltage $V_{DC2}^* = 320(V)$, the waveform of $i_L$ resembles that of the reference current $i_L^* = i_{DC}$.

When the motor is accelerated by a constant torque $T_e = 10(Nm)$ after 0.5(sec), we can observe that the stored power in $C_1$ is provided with the motor. Because the capacity of three phase power source is not enough, $V_{DC1}$ and $V_{DC2}$ decreases when the motor is accelerated.

Therefore, it is confirmed that the proposed control method is effective by both the simulation and the experiment.

5. Concluding remarks

This paper presents a novel control method of a regenerative power storage system composed by a bi-directional DC-DC converter and an electrolytic capacitor for electric machinery. We have designed a new control system using both a feed-back and a feed-forward path for a state-space averaged model of the bi-directional DC-DC converter. The theoretical stability analysis is also given. The effectiveness of the proposed control method is illustrated by both a simulation and an experiment.

References


Pattern Dynamics of Phase Synchronization in a Family of Coupled Several One-Dimensional Chaotic Maps

Masahiro Wada† and Yoshifumi Nishio‡

†Department of Information Science and Systems Engineering, Konan University
8–9–1 Okamoto, Higashinada-ku, Kobe 658-8501, Japan
‡Department of Electrical and Electronic Engineering, Tokushima University
2–1 Minami-Josanjima, Tokushima 770-8506, Japan.
Email: wada–m@konan-u.ac.jp, nishio@ee.tokushima-u.ac.jp

Abstract—In this study, phase synchronization behavior and control of its patterns in coupled chaotic maps are investigated. There are many types of chaotic map, then coupled chaotic systems yield wide variety of complex phenomena and further it is shown possibility to several engineering applications. The chaotic maps which have been governed by \( n \)-th power polynomial or sinusoidal functions is properly selected as a chaotic cell, then each chaotic map is connected to neighbors as a ring array or network structure. Several phase synchronization patterns and its control method are shown.

1. Introduction

We have now interests how to various patterns in nature were created. Coupled chaotic systems attract many researchers’ attention as a good model which can realize the complicated phenomena in the natural world, and further its dynamics can yield a wide variety of complex and strange phenomena. The coupled systems existing in nature exhibit great variety of phenomena such as complex mechanisms for all of the systems in the universe. These phenomena can be found in a metabolic network, a human society, the process of a life, self organization of neuron, a biological system, an ecological system and so many nonlinear systems. Among the studies on such coupled systems, many interesting researches relevant to the spatio-temporal chaos phenomena on the coupled chaotic systems have been studied until now, e.g. mathematical model in one- or two-dimensional network investigated earnestly by Kaneko [1]-[4], and found in physical circuit model [5]. The construction of multi-agent system on the coupled cubic map system has been reported [6]. Moreover, research of complicated phenomena and emergent property in the coupled cubic maps on 2-dimensional network system has been also reported [7]. The studies of coupled map lattice(CML), globally coupled maps(GCM) and so many studies concerned with such complex systems provided us tremendous interesting phenomena. This is an interesting report that phase dynamics are controlled due to change its parameter themselves. We had also reported the research on spatio-temporal phase patterns in coupled maps using a fifth-power function [8][9], in which it has been carried out in the unique case. However many coupled chaotic systems have wide variety of features and moreover its dynamics is also expected to be applied much engineering applications, there are many problems which should be solved in large scale coupled network systems by their complexity.

In this study, spatio-temporal chaotic behavior in coupled chaotic maps is investigated from the point of view in more faithful natural world. The chaotic map which has been governed by \( n \)-th power polynomial or sinusoidal functions is properly selected as a chaotic cell. We consider the model which chaotic cells are mutually connected to neighbors as a ring structure (i.e. CML type) by arbitrary coupling strength. Then, we show some phenomena which spatio-temporal chaos, complex behavior and several phase patterns can be found in the proposed coupled systems. Furthermore, its control dynamics are realized by changing a perturbation parameter.

2. Model Description

Chaotic maps are generally used for several approaches to investigate chaotic phenomena on coupled chaotic systems. Especially, the logistic map and the other types of chaotic maps such as a cut map, a circle map, a tent map, a cubic map are well known and popular. Obviously, it is necessary to have a lot of equilibrium points with the complex phenomena that corresponds to the natural world. Let us consider two types of chaotic map. Firstly, the chaotic map from an \( n \)-th power polynomial function written as follows.

\[
f(x) = \sum_{i=1}^{n} a_i x^i + \varphi \tag{1}\]

where \( a_i \) is a parameter which can determine for their chaotic feature, further \( \varphi \) is a new parameter for perturbation as a small variable value. The parameter \( \varphi \) should be normally set as zero. If it is needed to adopt the map with respect to the origin, odd-numbered coefficients \( a_i \) are only set suitable values in (1). In other words, even-numbered coefficients are set as all zero. Then, we can easily confirm that it generates chaos in this function. The some diagrams of the function (1) are shown in Fig. 1 with some equilibrium points.

Secondary, the chaotic map from a sinusoidal function written as follows,

\[
f(x) = e^{ix} \sin(ax) + bx + \varphi \tag{2}\]
where $a$, $b$ and $c$ are parameters which can determine for their chaotic feature. Especially the parameter $b$ is an important factor in order to suppress its divergence. The some diagrams of the function (2) are also shown in Fig. 1. From (1) and (2), it can be calculated rigorously several bifurcation conditions and boundary region.

In order to evaluate the function (1), Lyapunov exponent can be calculated as follows.

$$\lambda = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \log \left| \frac{df(x_k)}{dx} \right|$$

Lyapunov exponent is a very important measurement often used to show the existence of chaos. Some Lyapunov exponents with bifurcation diagram by changing one parameter are shown in Fig. 2. These are typical results which can be obtained from computer calculation. In case of using polynomial functions, period doubling and tangent bifurcation can be confirmed. On the other hand, in case of using sinusoidal functions, complicated bifurcation property can be confirmed even if the result of Lyapunov exponent is seen. Therefore chaotic maps possessing several equilibrium points can yield various wide interesting behavior.

3. Phase Patterns in Coupled Chaotic Maps

In this section, we consider a coupled chaotic system that one of these maps as a chaotic subsystem in each cell. It can be considered easily that coupled chaotic systems have wide variety of phase patterns. The term “spatio-temporal” is extensively used for irregular dynamical behavior observed from large scale complex systems of the relevant to both time and space. In this study, in order to confirm spatio-temporal chaos or phase patterns in the faithful natural world, consider a coupled model of the chaotic maps which are connected to neighbors on a ring array structure as shown in Fig. 3. Each chaotic cell is connected to neigh-
bors by arbitrary coupling strength $\varepsilon$. The total system by CML is represented as
\[
x_k(t+1) = (1-\varepsilon)f(x_k(t)) + \frac{\varepsilon}{2} \left( f(x_{k-1}(t)) + f(x_{k+1}(t)) \right),
\]
where $t$ is an iteration, $k$ is an index number of the cell which follows the cyclic rule, and $N$ is a size of coupled cell number, respectively.

Some numerical simulation results of model (4) for $N = 50$ are shown in Fig. 4 with coupling strength $\varepsilon = 0.30$. The initial condition for each cell is given as $x_k(0) \in [0.49, 0.51]$ uniformly. The figure indicates a grade of synchronization state for phase difference, with gray scale colors between white and black which correspond to synchronous and asynchronous state, respectively. Hereby the synchronous state with gray scale colors in 100 steps is displayed. A lot of interesting phenomena were confirmed though all the results can not be represented more here.

4. Pattern Dynamics and Control

Further, we attempt to control phase patterns of entire coupled system to become synchronous state with additional swing of function or changing the coupling strength. As shown in Fig. 4, it is confirmed that some parts are asynchronous state. Although all subsystem is the same, it is difficult to perform to control entire system synchronously. However control method should be simple as possible. Therefore, we propose two simple methods below.

(method 1)
\[
\varepsilon_k = \varepsilon^* \quad \text{if} \quad |x_k - x_{k+1}| < 0.5
\]

(method 2)
\[
\varepsilon_k = \varepsilon^* \quad \text{if} \quad |x_k - x_{k+1}| < 0.5
\]

\[
\varphi = \begin{cases} 
-0.1 & \text{if} \quad x_k \geq 0 \\
+0.1 & \text{if} \quad x_k < 0 
\end{cases}
\]

In case of method 1, it is changed only coupling parameter $\varepsilon_k$ in each cell. When the value of difference between two neighbors is larger than 0.5, the coupling strength $\varepsilon_k$ of the target cell changes to $\varepsilon^*$. In case of method 2, we use the condition of swing parameter $\varphi$ in addition to the method 1. Figure 5 shows a pattern obtained by method

Figure 4: Simulation results of phase synchronization state in coupled chaotic maps as a ring array for $N = 50$: (a) $n = 5$, $a_5 = 6.20$, $a_3 = -10.0$, $a_1 = 4.10$, (b) $n = 7$, $a_7 = -2.85$, $a_5 = 11.6$, $a_3 = -14.6$, $a_1 = 5.80$, (c) $n = 9$, $a_9 = 2.08$, $a_7 = -13.15$, $a_5 = 27.4$, $a_3 = -21.3$, $a_1 = 5.46$, and (d) $\varepsilon^{-|0.9|} \sin(10x) + 0.9$. Each result is performed in the condition $\varepsilon = 0.30$. 

Figure 3: Coupled chaotic system as a ring array or coupled lattice structure.
2 when the control is executed at $t = 200$. A lot of parts are made synchronous can be confirmed. To clarify a synchronous ratio, the difference of the adjoined cells counted the number of 0.2 or less and it was calculated. The mean values of the increase ratio before and after the control when the initial condition is changed and tries it 1000 times are shown in Table 1 and 2. The result that either was excellent was not obtained because two methods were carried out by a specific parameter, and the control method should be changed according to the characteristic of the chaotic map. Therefore, it is necessary to investigate whether other parameters are changed, and which control method is excellent though only control by the change in the coupling strength and the swing parameter were used in this experiment.

5. Conclusions

In this paper, some chaotic maps by $n$-th power polynomial or sinusoidal functions for using as a chaotic cell of coupled network have been proposed. Some illustrated computer simulation results of spatio-temporal chaotic behavior and several phase patterns in coupled chaotic maps have been shown. The simple method for controlling the patterns has been proposed, and the effectiveness has been investigated. We conclude that the supposed or similar coupled chaotic systems can be regarded as a good model for realizing complex phenomena in the universe concerned with self organization, mechanisms of pattern formation and so on. However some studies of pattern dynamics and the mechanism of clustering phenomena in such complex phenomena and many works have been left.

Table 1: Increase ratio of synchronous state when the setting parameters are the same as figure 4.

| $\epsilon^*$ | 0.4  | 0.5  | 0.6  | 0.7  | 0.8 
|----------------|-------|-------|-------|-------|-------
| $n = 5$        | 11.7  | 4.4   | 6.7   | 9.0   | 11.7 |
| $n = 7$        | 9.1   | 13.4  | 14.5  | 15.0  | 13.4 |
| $n = 9$        | 21.8  | 23.9  | 25.3  | 28.9  | 25.8 |
| $\epsilon^* \sin \phi$ | 6.4   | 6.0   | 2.5   | -6.1  | -12.1 |

Table 2: Increase ratio of synchronous state when the setting parameters are the same as figure 4 with swing parameter $\varphi$.

<table>
<thead>
<tr>
<th>$\epsilon^*$</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 5$</td>
<td>-8.1</td>
<td>-0.7</td>
<td>-0.9</td>
<td>0.8</td>
<td>4.3</td>
<td>2.7</td>
</tr>
<tr>
<td>$n = 7$</td>
<td>17.1</td>
<td>20.8</td>
<td>20.8</td>
<td>34.1</td>
<td>22.8</td>
<td>13.3</td>
</tr>
<tr>
<td>$n = 9$</td>
<td>8.0</td>
<td>16.3</td>
<td>17.3</td>
<td>17.6</td>
<td>18.7</td>
<td>15.6</td>
</tr>
<tr>
<td>$\epsilon^* \sin \phi$</td>
<td>-0.1</td>
<td>6.6</td>
<td>6.0</td>
<td>2.4</td>
<td>-6.1</td>
<td>-12.2</td>
</tr>
</tbody>
</table>

Acknowledgment

This work was partly supported by GRANT-IN-AID for "Open Research Center" project from MEXT, Japan.

References

Supertransients in coupled map lattice with distributed feedback control

Keiji Konishi

Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531, JAPAN
Email: konishi@eis.osakafu-u.ac.jp

Abstract—The present paper shows the supertransients in the coupled map lattice with distributed feedback control. The sufficient conditions for avoiding the supertransients are derived analytically. These conditions are numerically confirmed by the coupled logistic map lattice. In addition, the influence of the feedback gain onto the supertransients is investigated: we find that the dead beat feedback gain is not an optimal gain to shorten the transient time.

1. Introduction

Since coupled map lattice (CML) shows many interesting spatio-temporal nonlinear phenomena, it has been used as one of the most attractive systems in the field of nonlinear science [1]. When each site without coupling behaves periodically in chaotic region (i.e., periodic window), CML with weak coupling takes a long transient time to reach the periodic orbit. The transient shows a complicated spatio-temporal pattern. Kaneko found this phenomenon and called it supertransients [2, 3].

The controlling chaos has been the subject of many research papers and review articles [4, 5, 6, 7]. It is well known that the most fundamental method for controlling chaos is the OGY control method [8]. This method can provide a systematic procedure how to design the feedback controller on the basis of the modern control theory. The OGY control method for one-dimensional discrete-time systems is called occasional proportional feedback (OPF) method [9]. The several researchers have applied the OPF method to CML; however, to our knowledge, most of them focus on the local stability of the desired orbit [5, 6]. On the other hand, Youssefmir and Huberman reported that the distributed OPF method for CML takes a long transient time to reach the desired orbit [10]. If we want to use the desired orbit for engineering applications, the transient time should be shortened.

This paper shows that the supertransients can occur in the CML with the distributed OPF method. Furthermore, we derive the sufficient conditions for avoiding the supertransients. These conditions would be useful to design the CML and the controller. In addition, it is shown on numerical simulations that the dead beat feedback gain is not an optimal gain to shorten the transient time.

2. Supertransients

2.1. CML and its stability

This paper focuses on the coupled map lattice (CML) [1] described by

\[ x_{n+1}(i) = (1 - \varepsilon)f(x_n(i)) + \frac{\varepsilon}{2}(f(x_n(i-1)) + f(x_n(i+1))) \]  

(1)

for \( i = 1, 2, \ldots, N \). \( x_n(i) \in \mathbb{R} \) is the system state of site \( i \) at time \( n \). \( \varepsilon \in [0,1] \) is the coupling strength, \( f : [0,1] \to [0,\bar{x}] \) is the local nonlinear map, where \( \bar{x} \in [0,1] \) is the maximal value of \( f(x) \) for \( x \in [0,1] \). \( N \) is the number of sites. The boundary condition is periodic (i.e., \( x_n(0) = x_n(N), x_n(N+1) = x_n(1) \)). CML (1) has the uniform orbit,

\[ x_n(i) = x_n^* \text{ for } i = 1, 2, \ldots, N. \]  

(2)

This orbit implies that all the sites \( (i = 1, 2, \ldots, N) \) synchronize with the orbit \( x_n^* \) of the isolated map:

\[ x_{n+1}^* = f(x_n^*). \]  

(3)

It is known that CML (1) has the following two features [11].

\[ (\alpha) \text{ If the initial condition of all the sites } (i = 1, 2, \ldots, N) \text{ satisfies } 0 \leq x_0(i) \leq 1, \text{ then the site states } x_n(i) \text{ are bounded.} \]

\[ (\beta) \text{ The stability of the orbit } x_n^* \text{ of local map (3) is equivalent with that of uniform orbit (2) of CML.} \]

The feature (\( \alpha \)) guarantees that, if CML satisfies the above condition, then the site states \( x_n(i) \) are always in the region \([0,1]\) for any \( \varepsilon \in [0,1] \) and \( f : [0,1] \to [0,\bar{x}] \). The feature (\( \beta \)) shows that, if the local map (3) has the stable (unstable) fixed point \( x_0 = f(x_0) \), the uniform fixed point \( x_n(i) = x_0 \) is also stable (unstable) for any \( \varepsilon \in [0,1] \). This feature is valid for \( m \)-periodic orbits: \( x_n^* = f^m(x_n^*) \).

2.2. Supertransients in CML

This subsection numerically observes the supertransients found by Kaneko [2]. The logistic map is used as the local map \( f \):

\[ f(x) = ax(1 - x) : [0,1] \to [0,\bar{x}]. \]  

(4)
Figure 1: Supertransients of CML ($f(x) = ax(1-x)$, $a = 3.83$, $\varepsilon = 0.0012$, $N = 30$)

Figure 2: One-dimensional chaotic map without control $g(x_n, a^*)$ and dead-beat controlled map $g(x_n, h(x_n))$

Figure 3: Sketch of controlled CML

3. Supertransients in controlled CML

This section demonstrates that the supertransients can occur in CML controlled by distributed feedback, and provides the sufficient conditions to avoid it.

3.1. OPF method

Let us consider the one-dimensional system

$$x_{n+1} = g(x_n, a_n),$$

where $g(x, a) : [0, 1] \times R \rightarrow [0, \bar{x}]$ is a chaotic map. $\bar{x} \in [0, 1]$ is the maximal value of $g$. $a_n$ is the accessible system parameter which can be controlled every step within $|a_n - a^*| < d$, where $a^*$ is some nominal parameter value and $d \ll 1$ defines the range of parameter perturbation. For $a_n \equiv a^*$, system (5) behaves chaotically. The fixed point of system (5), $x_g$, is given by $x_g = g(x_g, a^*)$. In order to stabilize $x_n$ onto $x_g$, the OPF method uses the following control law:

$$a_n = h(x_n), \quad h(x) = \begin{cases} a^* & x \notin \Omega \\ k(x - x_g) + a^* & x \in \Omega \end{cases}$$

where $k \in R$ is the feedback gain. $\Omega := [x_{\min}, x_{\max}]$ denotes the neighborhood of $x_g$: $x_{\min} = x_g - |d/k|$, $x_{\max} = x_g + |d/k|$. If $k$ is chosen such that

$$|\eta + \gamma k| < 1,$$

then $x_g$ is stabilized, where $\eta = \{\partial g(x, a^*)/\partial x\}_{x=x_g}$, $\gamma = \{\partial g(x, a)/\partial a\}_{a=a^*}$. For simplicity, we use

$$k = k_{\text{opt}} := -\eta/\gamma$$

which realizes the dead beat control: $x_n \in \Omega$ reaches $x_g$ only by one step. Figure 2 sketches the chaotic map without control, $g(x_n, a^*)$, and the map controlled by the dead beat, $g(x_n, h(x_n))$.

3.2. CML controlled by distributed feedback

The controlled map $g(x_n, h(x_n))$ is used as the local map $f$ of CML (1) (see Fig. 3): $f(x_n(i)) := g(x_n(i), h(x_n(i)))$ ($i = 1, 2, \ldots, N$). As the controlled map $f$ satisfies the features $(\alpha)(\beta)$, then the uniform fixed point $x_n(i) = \bar{x}_g$ of the controlled CML is local stable for any $\varepsilon \in [0, 1]$. 

\[536\]
This section uses logistic map (4) as the local map \( g \): \( g(x,a) = ax(1-x) \). The fixed point \( x_g \) is given by \( x_g = (a^* - 1)/a^* \). The parameter values are set to \( a^* = 3.80 \), \( N = 30 \), \( d = 0.1 \). The feedback gain \( k \) is estimated by (8). The spatio-temporal diagram of controlled CML is shown in Fig. 4. If \( x_n(i) \in \Omega \), then the white square (\( \Box \) ) is plotted, otherwise (\( x_n \notin \Omega \) the black square (\( \blacksquare \) ) is plotted. These squares are plotted every 9 step. It can be seen that a long transient time is needed to reach the uniform fixed point \( x_n(i) = x_g \) (about 20,000 steps) and the transient pattern is complicated. It should be noted that this result is similar to the supertransients in Fig. 1. Therefore, we can say that the supertransients occur in the controlled CML.

### 3.3. Classification of transient states

Figure 5 shows the spatio-temporal diagrams of controlled CML for \( \varepsilon = 0.0130, \ 0.0298, \ 0.0400 \). The all parameters except for \( \varepsilon \) are the same as the numerical simulation in Fig. 4. We notice that Figs. 5 (a)~(c) have the following common feature: if \( x_n(i) \notin \Omega \) (\( \blacksquare \) ) , then there is no rule whether \( x_{n+1}(i) \) falls in \( \Omega \) (\( \Box \) ). On the contrary, we find that if \( x_n(i) \in \Omega(\Box) \), then \( x_{n+1}(i) \in \Omega(\Box) \) depends on the states of the next sites. We classify the dependence into three classes (I) (II) (II) below (see also Table 1).

**Class (I)** If \( x_n(i) \in \Omega(\Box) \), then \( x_{n+1}(i) \) falls in \( \Omega(\Box) \) independently of the next sites.

**Class (II)** If \( x_n(i) \in \Omega(\Box) \) and at least one next site falls in \( \Omega \) (i.e., \( x_{n}(i+1) \in \Omega \) or \( x_{n}(i-1) \in \Omega \), then \( x_{n+1}(i) \) falls in \( \Omega(\Box) \).

**Class (III)** If \( x_n(i) \in \Omega(\Box) \) and the next sites on both sides falls in \( \Omega \) (i.e., \( x_{n}(i+1) \in \Omega \) and \( x_{n}(i-1) \in \Omega \), then \( x_{n+1}(i) \) falls in \( \Omega(\Box) \).

### 3.4. Sufficient conditions to avoid supertransients

If the uniform orbit \( x_n(i) = x_n^* \) is desirable from an engineering point of view, a system designer would hope to shorten the amount of supertransient time or avoid supertransients. This subsection provides sufficient conditions to avoid it (The proof is omitted due to lack of space).

**Class (I)** If \( \varepsilon \) and \( d \) satisfies

\[
\varepsilon < |d\gamma/\eta|/\max\{\bar{x}-x_g, x_g\}, \quad (9)
\]

then class (I) occurs in the controlled CML with \( k = k_{opt} \).

**Class (II)** If \( \varepsilon \) and \( d \) satisfies

\[
\varepsilon < 2|d\gamma/\eta|/\max\{\bar{x}-x_g, x_g\}, \quad (10)
\]

then class (I) or (II) occurs in the controlled CML with \( k = k_{opt} \).

The parameters were set to \( d = 0.1, x_g = 2.8/3.8, \bar{x} = 3.8/4, \gamma = 0.1939, \eta = -1.8 \). Substituting these parameters into condition (9), we have \( \varepsilon < 0.0146 \). Figure 5 (a) (Class (I)) satisfies this condition. Furthermore, condition (10) is estimated as \( \varepsilon < 0.0292 \). Figure 5 (b) (Class (II)) satisfies this condition.

These conditions can allow the system designer to determine the coupling strength \( \varepsilon \) and the parameter perturbation range \( d \). If the supertransients in the controlled CML are undesirable, the system designer uses these parameters satisfying condition (9). When it is difficult to satisfy condition (9), the system designer can also avoid it by using condition (10).
### Table 1: Classification of transient states and their conditions

<table>
<thead>
<tr>
<th>Class</th>
<th>Space-time diagram in Fig. 5</th>
<th>Sufficient condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I)</td>
<td>(a)</td>
<td>$\varepsilon &lt; \frac{</td>
</tr>
<tr>
<td>(II)</td>
<td>(b)</td>
<td>$\varepsilon &lt; \frac{2</td>
</tr>
<tr>
<td>(III)</td>
<td>(c)</td>
<td>$\varepsilon &lt; \frac{3</td>
</tr>
</tbody>
</table>

Figure 6: Relation between the average transient time and gain $k$ ($a^* = 3.80, d = 0.1, N = 40, \varepsilon = 0.0360, 0.0390, 0.0410, 0.0420$)

### 3.5. Influence of $k$ onto supertransients

The previous subsections use the dead beat gain $k = k_{opt}$ given by (8). This subsection investigates the influence of $k$ onto the transient time. The relation between the average transient time and gain $k$ for $\varepsilon = 0.0360, 0.0390, 0.0410, 0.0420$ is numerically investigated as shown in Fig. 6. $k_{min}$ and $k_{max}$ are the minimum and maximum values satisfying stability condition (7). The average time is estimated from random 500 initial states. It can be seen that the gain width for the short transient decreases with increasing $\varepsilon$. Furthermore, all the lines in Fig. 6 show a minimum where the gain $k$ is smaller than the dead beat gain $k_{opt}$.

### 4. Conclusion

This paper shows that the supertransients can occur in the CML controlled by the distributed feedback (i.e. OPF method). The sufficient conditions for classes (I) and (II) are derived analytically. These conditions can be used to avoid the supertransients in designing the coupling strength and the parameter perturbation range. Furthermore, we show that the dead beat feedback gain is not an optimal gain to shorten the transient state.

The transient state in the one-way CML with the distributed delayed feedback control was mentioned in [12]. It is interesting to show whether our analysis in this paper can be applied to the CML with the distributed delayed feedback control.

### References


Stochastic Resonance in Coupled Costas Loops with Imperfect Phase Detectors

Hisato Fujisaka, Asuka Watanabe and Takeshi Kamio

Faculty of Information Sciences, Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194, Japan
Email: fujisaka@im.hiroshima-cu.ac.jp

Abstract—Single and coupled first-order Costas loops are considered as systems with spatially periodic potentials defined on one and N-dimensional tori. Imperfection of phase detectors in the loops adds spatially and temporally periodic fluctuations to the potentials. This paper shows by means of the linear response theory and numerical analysis that the fluctuations and additive noise cause stochastic resonance in the single and coupled loops.

1. Introduction

Stochastic resonance (SR) is one of interesting phenomena observed in nonlinear stochastic systems. Since active electronic circuits with noise sources are nonlinear stochastic systems, SR is observed in electronic circuits, for example, bi-stable circuits and threshold circuits [1]. However, the phenomena in single and coupled synchronization loops have not been studied thoroughly in spite of their importance in communication systems. We merely have a few reports on the study [2]. In this paper, we investigate whether or not single and coupled first-order Costas loops operate in SR condition. The loops are spatially periodic potential systems defined on one and N-dimensional tori $T^{1}, T^{N}$. Imperfection of phase detectors in the loops adds spatially and temporally periodic small fluctuations to the potentials. On the assumption that the fluctuations are small, we apply the linear response theory (LRT) [3] to investigate SR in the loops.

2. Coupled Synchronization Loops

We build a system of coupled synchronization loops. In the system, $N$(integer) Costas loops tracking QPSK carrier are arranged circularly. A loop given index $i(\in \{1, 2, \ldots, N\})$ connects with $2L(<N)$ loops with indices $i \pm k$, $k=1,2,\ldots,L$. The loops do not possess loop filters. Figure 1 shows synchronization loops with the nearest neighbor couplings.

Let a QPSK modulated carrier inputted to the loops be given by $\exp(I(\omega_0 t + \theta_N(t)))$, $I^2=-1$. Phase $\theta_{IN}(t)$ is modulated according to

$$e^{-I\theta_{IN}(t)} = \frac{1+I}{2} \text{sign}(\cos(\omega_m t)) + \frac{1-I}{2} \text{sign}(\sin(\omega_m t)) \tag{1}$$

The loops contain white noise sources $\sqrt{D}\Gamma_i(t)$ which are mutually independent,

$$<\Gamma_i(t)\Gamma_j(t)> = \delta_{ij} \tag{2}$$

Let a waveform generated by a voltage controlled oscillator (VCO) in each loop be expressed by $\exp(I(\omega_0 t + \theta_i(t)))$. Phase $\theta_i(t)(\in T=[0, 2\pi])$ is controlled by signal $v_i$ according to

$$\frac{d\theta_i(t)}{dt} = v_i \tag{3}$$

Phase detectors PDI are used for mutual synchronization between the loops. Output from PDI connecting loops with indices $i$ and $j$ is given by

$$u_{ij} = -\sin(\theta_i(t) - \theta_j(t)) \tag{4}$$
Phase detector PDII makes VCO track the QPSK carrier. The output from PDII is given by

\[ u_{i,IN} = -\sin(M(\theta_i(t) - \theta_{IN}(t))) - d\sin(\theta_i(t) - \theta_{IN}(t)) \]  

(5)

For QPSK carrier tracking, \( M=4 \). The second term with small parameter \( d \) represents imperfection of PDII (\( d=0 \) if PDII is ideal). The term is expanded to

\[ \sin(\theta_i(t) - \theta_{IN}(t)) = \sin(\theta_i(t))f_1(t) + \cos(\theta_i(t))f_2(t), \]

(6)

where \( f_1(t) \) and \( f_2(t) \) denote the real and the imaginary parts of \(-I\cdot\theta_{IN}(t)\) in Eq.(1). Outputs \( u_{i,j} \) and \( u_{i,IN} \) from PDI and PDII are multiplied by \( \epsilon \) and \( A_o \), and inputted to the VCO as control signals. Phase equation for each loop is then given by

\[ \frac{d\theta_i}{dt} = -A_o\sin(M\theta_i) \]

(7)

\[ -\epsilon \sum_{k=1}^{L} \{ \sin(\theta_i - \theta_{i-k}) + \sin(\theta_i - \theta_{i+k}) \} \]

\[ -A_m \{ \sin(\theta_i)f_1(t) + \cos(\theta_i)f_2(t) \} \]

\[ +\sqrt{D}\Gamma_i(t), \]

\[ A_m = dA_o, \quad i = 1, 2, \ldots, N \]

3. The Fokker-Planck Equation

Let \( \Theta \) denote a phase vector whose \( i \)-th entry is phase \( \theta_i \) of VCO,

\[ \Theta = (\theta_1, \theta_2, \ldots, \theta_N) \in \mathbb{T}^N \]

(8)

Probability distribution \( W(\Theta, t) \) of \( \Theta \) satisfies the following Fokker-Planck equation (FPE):

\[ \frac{\partial W(\Theta, t)}{\partial t} = L_{FP}W(\Theta, t) \]

\[ = \{L_o + \epsilon L_{cpl} + L_{PM}\}W(\Theta, t) \]

(9)

\[ L_o = \sum_{i=1}^{N} \{ A_o \frac{\partial}{\partial \theta_i} \sin(M\theta_i) + D\frac{\partial^2}{\partial \theta_i^2} \} \]

(10)

\[ L_{cpl} = \sum_{i=1}^{N} \frac{\partial}{\partial \theta_i} \{ \sum_{k=1}^{L} \{ \sin(\theta_i - \theta_{i-k}) + \sin(\theta_i - \theta_{i+k}) \} \} \]

(11)

\[ L_{PM} = A_m \{ f_1(t) \sum_{i=1}^{N} \frac{\partial}{\partial \theta_i} \sin(\theta_i) + f_2(t) \sum_{i=1}^{N} \frac{\partial}{\partial \theta_i} \cos(\theta_i) \} \]

(12)

In the rest of this section, we assume that phase detectors PDII are ideal, that is \( L_{PM} \equiv 0 \). From the phase equation (7) we recognize the coupled loops as a potential system with the following potential \( U(\Theta) \):

\[ U(\Theta) = -\frac{1}{D} \sum_{i=1}^{N} A_o \sin(M\theta_i) + \epsilon \sum_{i=1}^{N} \sum_{k=1}^{L} \sin(\theta_i - \theta_{i+k}) \]

(13)

The stationary solution and the probability current of FPE (9) are expressed by using the potential as follows:

\[ W_{st}(\Theta) = C_{st}e^{-U(\Theta)} \]

(14)

\[ S_i = -De^{-U(\Theta)} \frac{\partial}{\partial \theta_i} e^{U(\Theta)}W \]

(15)

Coefficient \( C_{st} \) is chosen such that \( W_{st} \) is normalized. From Eqs.(14) and (15) we find that the probability current is zero at the stationary condition, \( S_{i,st}=0 \).

We define an operator \( L \) as

\[ L = \epsilon \frac{2}{L} (L_o + \epsilon L_{cpl}) e^{-\frac{U}{2}} = \sum_{i=1}^{N} a_i^+ a_i \]

(16)

\[ a = \sqrt{De^{-\frac{U}{2}}} \frac{\partial}{\partial \theta_i} \sqrt{De^{-\frac{U}{2}}} \]

Since operators \( a \) and \( a^+ \) are adjoint, \( L \) is an Hermitian operator. Eigenfunctions \( \psi_n(\Theta) \) and \( \phi_n(\Theta) \) of operators \( L_o + \epsilon L_{cpl} \) and \( L \) are real and have a relation \( \psi_n(\Theta) = \exp(\sqrt{U/2})\phi_n(\Theta) \). Then, the two operators have the same eigenvalues \( \lambda_n \). We define inner product between two functions \( f(\Theta) \) and \( g(\Theta) \) as

\[ \langle f(\Theta), g(\Theta) \rangle = \int_{\mathbb{T}^N} f(\Theta)g(\Theta)d\Theta \]

(17)

We easily find that \( \lambda_n \leq 0 \) from

\[ \langle L\psi_n(\Theta), \psi_n(\Theta) \rangle = -\sum_{i=1}^{N} \langle a_i^+ \psi_n(\Theta), a_i \psi_n(\Theta) \rangle \]

(18)

and

\[ \langle L\psi_n(\Theta), \psi_n(\Theta) \rangle = \lambda_n ||\psi_n(\Theta)||^2 \]

(19)

Eigenfunctions \( \psi_n(\Theta) \) and \( \phi_n(\Theta) \) have orthonormality and completeness relations expressed by [4]

\[ \langle \psi_n(\Theta), \psi_m(\Theta) \rangle = \delta_{n,m} \]

(20)

and

\[ \delta(\Theta - \Theta') = \sum_{n=0}^{\infty} \psi_n(\Theta)\phi_n(\Theta') \]

(21)

4. Correlation Functions

Let the phase vector at time \( t' \) and \( t=t'+\tau \) be denoted by \( \Theta' \) and \( \Theta \). Autocorrelation function \( K_{ss}(\tau) \) of \( \sum_i \sin(\theta_i) \) is defined as

\[ K_{ss}(\tau) = \langle \sum_{i=1}^{N} \sin(\theta_i), \sum_{i=1}^{N} \sin(\theta_i') \rangle > 0 \]
\[ P(\Theta, t' + \tau|\Theta', t') = e^{(L_o + \epsilon L_{cpl})\tau} \delta(\Theta - \Theta') \]  
\hspace{1cm} (23)

Using Eqs. (14), (23) and the expression of completeness (21) we obtain
\[ K_{ss}(\tau) = C_{st} \sum_n \left\{ \int_{T^N} \left\{ \sum_{i=1}^{N} \sin(\theta_i) \right\} \varphi_n(\Theta)d\Theta \right\}^2 e^{\lambda_n \tau} \]  
\hspace{1cm} (24)

Substituting Eq. (23) for transition matrix \( P \) in Eq. (22) and differentiating it with respect to \( \tau \), we obtain
\[ \frac{dK_{ss}(\tau)}{d\tau} = \int_{T^N} \left\{ \sum_{i=1}^{N} \sin(\theta_i) \right\} e^{(L_o + \epsilon L_{cpl})\tau} \]  
\hspace{1cm} (25)

The following correlation function
\[ K_{sc}(\tau) = < \sum_{i=1}^{N} \sin(\theta_i), \sum_{i=1}^{N} \cos(\theta'_i) > \]  
\hspace{1cm} (26)

and its derivative can be obtained similarly as we derived Eqs. (24) and (25).

5. Linear Response Theory

When phase detectors PDI are imperfect, probability distribution \( W \) contains a small time dependent component \( w(\Theta, t) \). Substituting \( W_{st} + w \) for \( W \) in FPE (9) and neglecting \( L_{PM}w \), we obtain
\[ \frac{\partial w}{\partial t} = (L_o + \epsilon L_{cpl}) w + L_{PM}W_{st} \]  
\hspace{1cm} (27)

A formal solution of this equation is
\[ w = \int_{-\infty}^{t'} e^{(L_o + \epsilon L_{cpl})(t-t')} L_{PM}W_{st} dt' \]  
\hspace{1cm} (28)

Deviation \( \Delta_{\sin}(t) \) of the expectation of \( \sum_i \sin(\theta_i) \) is then given by
\[ \Delta_{\sin}(t) = \int_{T^N} \sum_{i=1}^{N} \sin(\theta_i) w d\Theta \]  
\hspace{1cm} (29)

\[ \Delta_{\sin}(t) = A_m \left( \int_{-\infty}^{\infty} R_{sc}(t-t') f_2(t') dt' \right) \]  
\hspace{1cm} (30)

Response functions \( R_{ss} \) and \( R_{sc} \) should be zero for \( \tau < 0 \). The deviation is given in the frequency domain by
\[ \tilde{\Delta}_{\sin}(\omega) = A_m \left\{ \tilde{\chi}_{ss}(\omega) F_1(\omega) + \tilde{\chi}_{sc}(\omega) F_2(\omega) \right\} \]  
\hspace{1cm} (32)

Then, we obtain a component \( \tilde{\Delta}_{\sin}(\omega_m) \) of \( \Delta_{\sin}(t) \) at modulation frequency \( \omega_m \). Operating \( L_o + \epsilon L_{cpl} \) on \( \sum_i \sin(\theta_i)W_{st} \) and using \( (L_o + \epsilon L_{cpl})W_{st}=0 \) and \( S_{st}=0 \), we derive the following equation:
\[ (L_o + \epsilon L_{cpl})(\sum_{i=1}^{N} \sin(\theta_i)W_{st}) = D \sum_{i=1}^{N} \frac{\partial}{\partial \theta_i}(\cos(\theta_i)W_{st}) \]  
\hspace{1cm} (33)

From Eqs. (25), (31) and (33) we obtain the following relation (34), and similarly Eq. (35).
\[ R_{sc}(\tau) = \frac{1}{D} \frac{dK_{ss}(\tau)}{d\tau} \]  
\hspace{1cm} (34)

\[ R_{ss}(\tau) = -\frac{1}{D} \frac{dK_{sc}(\tau)}{d\tau} \]  
\hspace{1cm} (35)

6. Signal-to-Noise Ratio

After computing the eigenvalues and the eigenfunctions of operator \( L_o + \epsilon L_{cpl} \) and substituting them in Eq. (24), we obtain correlation function \( K_{ss}(\tau) \). We also obtain \( K_{sc}(\tau) \) as it also is expressed with
we obtain a component $\tilde{\Delta}$ shown in Eq. (32). According to the similar procedure we obtain a component $\tilde{\Delta}_{cos}(\omega_m)$ of $\sum_i \cos(\theta_i)$. We consider the two components as signal components.

When phase detectors PDII are ideal, we obtain the power spectral density (PSD) of $\sum_i \sin(\theta_i)$ by the Fourier transform of $K_{ss}(\omega)$, for which $K_{ss}(\omega)$ stands. We can also obtain the Fourier transformed function $K_{cc}(\omega)$ for $\sum_i \cos(\theta_i)$ similarly. We assumed in the previous section that deviation $w(\Theta, t)$ of probability distribution $W(\Theta, t)$ from stationary distribution $W_{st}(\Theta)$ is small. Thus, $\tilde{K}_{ss}(\omega)$ and $\tilde{K}_{cc}(\omega)$ approximate PSD of the noise components of $\sum_i \sin(\theta_i)$ and $\sum_i \cos(\theta_i)$.

We then define signal-to-noise ratio (SNR) for the coupled loops as

$$\text{SNR} : 10 \log \frac{\tilde{\Delta}_{sin}(\omega_m)^2 + \tilde{\Delta}_{cos}(\omega_m)^2}{K_{ss}(\omega_m) + K_{cc}(\omega_m)} \quad (36)$$

Figure 2 shows internal noise power $D$ versus SNR curves. Coupled loops are constructed on the condition that $L=1$ (the nearest neighbor coupling), $A_o = 1.0, A_m = 0.3, \epsilon = 0.15, \omega_m = 25\pi/2^{11}$. In LRT analysis, we neglected high frequency components $f_1(k\omega_m)$ and $f_2(k\omega_m), k \geq 2$, of modulation signals $f_1(t)$ and $f_2(t)$. When noise power $D$ is in the range of Fig. 2, a term composed of eigenvalue $\lambda_1$ whose magnitude is the smallest and corresponding eigenfunction $\varphi_1$ is dominant in Eq. (24). Thus, we neglected terms composed of $\lambda_n$ and $\varphi_n, n \geq 2$. In numerical simulation, we substituted the averaged power of the spectrum of $\sum_i \sin(\theta_i)$ and $\sum_i \cos(\theta_i)$ in an interval $[0.8\omega_m, 1.2\omega_m]$ for $K_{ss}(\omega_m)$ and $K_{cc}(\omega_m)$ in Eq. (36) and computed the SNR.

Let eigenvalues and eigenfunctions of the Fokker-Planck operator for a single loop be denoted by $\lambda_{scl,n}$ and $\varphi_{scl,n}(\theta)$. We approximated $\varphi_{scl,n}(\theta)$ in the form of truncated Fourier series. When the coupling strength is zero, $\epsilon = 0$, eigenvalues $\lambda_{cpl,n}^{(0)}$ and eigenfunctions $\varphi_{cpl,n}^{(0)}(\Theta, t)$ of the Fokker-Planck operator for the coupled loops are given by the sums of $\lambda_{scl,n}$ and the products of $\varphi_{scl,n}(\theta)$. For $\epsilon \neq 0$, approximated eigenvalues $\lambda_{cpl,n}^{(1)}$ and eigenfunctions $\varphi_{cpl,n}^{(1)}(\Theta, t)$ by adding perturbation terms to $\lambda_{cpl,n}^{(0)}$ and $\varphi_{cpl,n}^{(0)}(\Theta, t)$. For the perturbation method to be applicable, parameter $\epsilon$ must be small, namely, the loops must couple weakly with one another.

7. Conclusions

We have indicated by LRT analysis and verified by numerical simulation that a Costas loop with an imperfect phase detector for QPSK carrier tracking and its coupled loops can be in SR conditions. We will be able to show also that single and coupled synchronization loops for tracking BPSK and other phase modulated carriers are in SR condition in the same way.

The correlation functions derived in section 4 will be useful for investigating noise-induced synchronization phenomena in the coupled loops. The investigation is one of our future subjects.

References


Asymmetrical Chaotic Coupled System Using Two Different Parameter Sets.

Yasuteru Hosokawa† and Yoshifumi Nishio‡

†Department of Information Science, Shikoku University
Furukawa, Ohjin-cho, Tokushima 771-1192 Japan
‡Department of Electrical and Electronics Engineering, Tokushima University
2-1 Minami-Johsanjima-cho, Tokushima 770-8506 Japan
Email: hosokawa@keiei.shikoku-u.ac.jp†, nishio@ee.tokushima-u.ac.jp‡

Abstract—In our previous study, we observed the interesting phenomenon in an asymmetrical chaotic coupled system. It is that the synchronous rate of the one subcircuits group increases in spite of increasing parameter mismatches in the system.

In this study, in order to verify this phenomenon using another system, an asymmetrical chaotic coupled system is proposed and investigated. Asymmetry of the system is realized by using two parameter sets. In the case of five subcircuits, we confirmed synchronization phenomena in computer calculations. Additionally, it was confirmed that synchronous rates of subcircuits using one parameter set are increased by increasing a parameter mismatch rate of the other subcircuits. We consider that this result is corresponding to results of previous study.

1. Introduction

Coupled systems of chaotic subsystems generate various kinds of complex higher-dimensional phenomena such as spatio-temporal chaotic phenomena, clustering phenomena, and so on. One of the most studied systems may be the coupled map lattice proposed by Kaneko[1]. The advantage of the coupled map lattice is its simplicity. However, many of nonlinear phenomena generated in nature would be not so simple. Therefore, it is important to investigate the complex phenomena observed in natural physical systems such as electric circuits systems[2]-[6].

In our previous study[7], synchronization phenomena in an asymmetrical coupled system of chaotic circuits were investigated. The system is coupled globally and the coupling elements are resistors. Each subcircuit has two coupling nodes and the asymmetrical coupled system is realized by selecting one of two coupling nodes. This system was investigated as non-ideal system. The small parameter mismatches were given to the subcircuits as a mismatch of the oscillation frequency. As a result of investigating this system, interesting phenomenon was observed. The phenomenon is that the synchronous rate of the one subcircuits group increases in spite of increasing parameter mismatches in the system. We suppose that the phenomenon can be explained as follows. The synchronizations of the one subcircuits group and other subcircuits group are constricted each other. Therefore, in the case of decreasing the synchronization of one group decreases, the synchronization of the other group increases.

In this study, in order to verify this phenomenon using another system, an asymmetrical chaotic coupled system is proposed and investigated. Asymmetry of the system is realized by using two parameter sets. In the case of five subcircuits, we confirm synchronization phenomena in computer calculations. Additionally, it is confirmed that synchronous rates of subcircuits using one parameter set are increased by increasing a parameter mismatch rate of another subcircuits.

2. Asymmetrical Coupled Systems

Figure 1: Asymmetrical coupled chaotic system.

Figure 2: Chaotic subcircuit.
ments are resistors as shown in Fig. 1. The coupled subcircuit is a chaotic circuit[8] as shown in Fig. 2. In our previous study, asymmetry is realized as selecting one of two nodes in Fig. 2. In this study, only one node are used and asymmetry is realized by using two parameter sets. The circuit using one parameter set is called as "A-subcircuits", and using the other parameter sets is called as "B-subcircuit". Now, in order to carry out computer calculation, circuit equations are derived. Figure 3 shows a nonlinear resistor which consist of two diodes. Circuit equations are derived using Figure 3(b). We define m and n as the number of circuits. The number of A-subcircuits is m, and the number of B-subcircuits is n.

\begin{align}
C_{1a} \frac{dv_{k1}}{dt} &= g v_{k1} - G_d f(v_{k1} - v_{k2}) + G \left\{ \sum_{i=1}^{m+n} v_i - (m+n)v_{k1} \right\}, \\
C_{2a} \frac{dv_{k2}}{dt} &= -i_{k3} + G_d f(v_{k1} - v_{k2}), \\
L_a \frac{di_{k3}}{dt} &= (1 + p_k) v_{k2},
\end{align}

\tag{1}

B-subcircuits(m ≤ k ≤ m + n):

\begin{align}
C_{1b} \frac{dv_{k1}}{dt} &= g v_{k1} - G_d f(v_{k1} - v_{k2}) + G \left\{ \sum_{i=1}^{m+n} v_i - (m+n)v_{k1} \right\}, \\
C_{2b} \frac{dv_{k2}}{dt} &= -i_{k3} + G_d f(v_{k1} - v_{k2}), \\
L_b \frac{di_{k3}}{dt} &= (1 + q_k) v_{k2},
\end{align}

\tag{2}

where,

\[
f(v) = v + \frac{|v - V_{th}|}{2} - \frac{|v + V_{th}|}{2}.
\]

Function f(v) is characteristics of the nonlinear resistor shown in Fig. 3. Parameters p_k and q_k shows small parameter mismatches. Using the following parameters and variables,

\[
x_k = \frac{v_{k1}}{V_{th}}, \quad y_k = \frac{v_{k2}}{V_{th}}, \quad z_k = \frac{1}{V_{th}} \sqrt{\frac{L_a}{C_{2a}}}, \\
t = \sqrt{\frac{L_a C_{2a}}{C_{1a}}} t, \quad \alpha = \frac{C_{2a}}{C_{1a}}, \\
\beta = g \sqrt{\frac{L_a}{C_{2a}}} \quad \gamma = G_d \sqrt{\frac{L_a}{C_{2a}}}, \quad \delta = G \sqrt{\frac{L_a}{C_{2a}}}, \\
\varepsilon = \frac{C_{2a}}{C_{1b}} \quad \zeta = \frac{C_{2a}}{C_{2b}} \quad \text{and} \quad \eta = \frac{L_a}{L_b},
\]

\tag{3}

Normalized equations are described as follows.

A-subcircuits(1 ≤ k ≤ m):

\[
\begin{align*}
\dot{x}_k &= \alpha \beta x_k - \alpha \gamma f'(x_k - y_k) \\
&\quad + \alpha \delta \left\{ \sum_{i=1}^{m+n} x_i - (m+n)x_k \right\}, \\
\dot{y}_k &= -z_k + \gamma f'(x_k - y_k), \\
\dot{z}_k &= (1 + p_k) y_k,
\end{align*}
\]

\tag{4}

B-subcircuits(m + 1 ≤ k ≤ m + n):

\[
\begin{align*}
\dot{x}_k &= \varepsilon \beta x_k - \varepsilon \gamma f'(x_k - y_k) \\
&\quad + \varepsilon \delta \left\{ \sum_{i=1}^{m+n} x_i - (m+n)x_k \right\}, \\
\dot{y}_k &= \zeta \left\{ -z_k + \gamma f'(x_k - y_k) \right\}, \\
\dot{z}_k &= \eta (1 + q_k) y_k,
\end{align*}
\]

\tag{5}

where,

\[
f'(x) = x + \frac{|x - 1| - |x + 1|}{2}.
\]

We carry out computer calculations in the case of five subcircuits. The system consists of two A-subcircuits and three B-subcircuits. Figures 4 and Figure 5 show computer calculated results using following initial values and parameters. Figure 4(a) show a attractor of A-subcircuit (k = 1), Figure 4(b) show a attractor of B-subcircuit (k = 3). Horizontal axes are x_k and vertical axes are z_k. Double scroll type attractors are observed. Figures 5 show voltage differences between each subcircuits. Vertical axes show voltage differences and horizontal axes show time. Namely, in the case of synchronizing two subcircuits, the amplitude becomes zero. Value Q is corresponding to parameter mismatch rate of B-subcircuits.
The first graph shows the voltage difference between two subcircuits. In the case of Eq. (6) and $Q = 0.05$.

$$x_k(0) = y_k(0) = z_k(0) = 0.100, \quad (k = 1, 2, 3, 4, 5),$$
$$m = 2, \quad n = 3, \quad \alpha = 0.600, \quad \beta = 0.500, \quad \gamma = 20.0,$$
$$\delta = 0.070, \quad \varepsilon = 0.6, \quad \zeta = 1.5, \quad \eta = 0.5$$

$$p_k = 0.001 (k - 1) \text{ and } q_k = Q(k - 1).$$

The second graph shows the voltage difference between an A-subcircuit and a B-subcircuit. These are not synchronized at all. The third and fourth graphs show the voltage differences between two B-subcircuits. These results are similar than first graph. However these are not the same. In the case of previous study, corresponding results are shown in Fig. 6. From comparing two results, we can see as follows: Fig. 5(a) is similar to the case of lower $Q$ value than Fig. 6. Fig. 5(b) is similar to Fig. 6. In Fig. 5(c), synchronous rate is higher than Fig. 6. Now, in order to investigate the relationship between synchronous rates and parameter mismatches, we define the synchronization as shown in Fig. 7 and follows:

$$|x_k - x_{k+1}| < 0.01$$

where $k$ is number of subcircuits.

Fig. 8 shows synchronous rate averages between two subcircuits to $Q$ which is corresponding to parameter mismatch rate. The synchronous rate is defined as a ratio of synchronous time and total time. Fig. 9 shows a result of previous study corresponding to Fig. 8. A-node of Fig. 9 is corresponding to A-subcircuit, and B-node is corresponding to B-subcircuit. In Fig. 8 and Fig. 9, the synchronous rates of one group are increasing by increasing $Q$. We can consider to obtain the same phenomena. Additionally, thirty subcircuits case is investigated. Figures 10 and 11 shows the result. We can also observed the same result. Namely, in spite of increasing parameter mismatch rate, synchronous rate of A-subcircuit increasing.

3. Conclusions

In this study, in order to verifier the phenomenon of previous study, an asymmetrical coupled system is proposed and investigated. Asymmetry of the system is realized by using two parameter sets.

In the case of five subcircuits, we confirmed similar phenomena in computer calculations. Additionally, It was confirmed that synchronous rates between circuits using one parameter set are increased by increasing a parameter mismatch rate of the other parameter set. We consider that this
Figure 8: Synchronous rate of chaotic circuits to the parameter mismatch of B-subcircuit in the case of five subcircuits. $m = 2$, $n = 3$, $\alpha = 0.600$, $\beta = 0.500$, $\gamma = 20.0$, $\delta = 0.070$, $\varepsilon = 0.6$, $\zeta = 1.5$ and $\eta = 0.5$.

Figure 9: Synchronous rate of previous study corresponding to the parameter mismatch of B-subcircuit in the case of five subcircuits. $m = 2$, $n = 3$, $\alpha = 0.400$, $\beta = 0.500$, $\gamma = 20.0$, and $\delta = 0.07$.

The result is corresponding to results of previous study. Furthermore, we also confirmed the same phenomena in the case of thirty subcircuits. Therefore, we consider that these phenomena are not the phenomena generated on a special system.

References


Parasitic capacitances and chaotic dynamics in a log-domain oscillator

Alon Ascoli*, Orla Feely*, and Paul Curran*

* School of Electrical, Electronic, and Mechanical Engineering, University College Dublin, Belfield, Dublin 4, Ireland
E-mails: alon.ascoli@hotmail.com, orla.feely@ucd.ie, paul.curran@ucd.ie

Abstract—This paper presents an analytical model of the chaotic dynamics of a recently proposed log-domain oscillator. It is shown how the inclusion of parasitic capacitances on the Ebers-Moll large-signal static model of each bipolar junction transistor permits the derivation of nonlinear differential equations that capture the chaotic behaviour of the oscillator.

1. Introduction

In log-domain current-mode circuits [1] the large-signal static forward-active mode exponential current-voltage relationship of the bipolar junction transistor (BJT) is used first to compress the input currents to the logarithmic domain, where analog processing occurs, and then to expand the resulting output voltages back to the linear domain in order to restore input-output linearity. Due to the pre-compressing and post-expanding operations, log-domain circuits exhibit a number of desirable features for integrated circuit (IC) realizations, such as low distortion levels, high signal-to-noise ratios, wide dynamic range, small power consumption, fine speed, and large bandwidth. Moreover, these circuits show ease of electronic tuneability of the key frequency parameters.

Log-domain chaotic oscillators are commonly designed by applying the exponential state-space transformation to a known set of nonlinear differential equations, which displays chaos under special conditions [2-3]. If it is assumed that each BJT employed in the final oscillator introduces a large-signal static exponential nonlinearity, as is common practice in log-domain circuit design [1], the mapped system may then be interpreted as a set of nodal equations, which are finally realized by using grounded capacitors and constant current sources, positive and negative log-domain cells [1], and a sub-circuit implementing the particular nonlinearity present in the original system. However, in [4] it was noted that, due to the limited range of the exponential function, if, as is usually the case, the positive orthant is not an invariant set for the original system, the above mentioned state-space transformation is not a diffeomorphism, because it is not even defined in the negative orthant. It follows that the original system and the mapped one are not dynamically equivalent. Nonetheless, taking the state-space transformation in a formal sense only, and then assuming the above mentioned BJT static model, a log-domain oscillator may still be devised as a formal transistor-level realization of the mapped system. In [4] this two-step formal procedure was applied to an original system with tanh nonlinearity, and led to a novel physically realizable log-domain oscillator. Surprisingly, however, the dynamics of this oscillator are distinct from those of the mapped system, and were therefore thoroughly investigated in [4] with tools from the theory of nonlinear dynamics [5]. Because of the previously mentioned attractive properties of log-domain circuits, the use of the novel log-domain oscillator from [4] may be appropriate in chaos-based communication systems. However, a valid analytical model for the chaotic dynamics of the oscillator, not presented so far, is needed.

The studies which led to this paper were focused on the derivation of the actual nonlinear differential equations describing the chaotic behaviour of the oscillator. Due to the internal nonlinearity of log-domain circuits [6], their dynamics may be strongly influenced by the tiny BJT parasitic capacitances [7]. The log-domain oscillator from [4] is not an exception: each transistor employed in the oscillator introduces a large-signal dynamic nonlinearity, which may be tracked only by including BJT internal capacitances for modeling purposes. This is the reason behind the failure of the mapped system, based upon the BJT static model, to explain the dynamics of the oscillator. In this paper the simplest transistor dynamic model is first used to derive an analytical model capturing the dynamics of the oscillator from [4], and then validated by comparing circuit simulations with numerical integrations of the proposed equations.

2. Log-domain Chaotic Oscillator

2.1. Circuit Design Review

The novel chaotic log-domain oscillator proposed in [4] originates from application of a two-step formal procedure to the following nonlinear differential equation:

\[ \dot{x} = -ax - x + 2 \tanh(x). \tag{1} \]

Equation (1), in which differentiation is with respect to a normalized time \( \tau \), displays chaos for \( a = 0.19 \). Letting \( x = x_3 \), a canonical state-space representation of (1) is:

\[
\begin{align*}
\dot{x}_1 &= \omega_o (x_3 - 2 \tanh(x_3 - m') + m') \\
\dot{x}_2 &= \omega_o (x_1 - x_3) \\
\dot{x}_3 &= \omega_o (x_2 - ax_3)
\end{align*}
\]  

(2)
where differentiation is with respect to time $t = \omega \tau'$ ($\omega$ is a positive normalizing factor), and $m'$ is a positive constant which ensures that all the state variables $x_i$ ($i \in \{1, 2, 3\}$) are positive at equilibrium. Since the positive orthant is not an invariant set for (2), then the following exponential state-space transformation
\[
x_i = n_i I'_2(C_S V_\text{CT})^2 \exp(V_i V_\text{T}) \text{ for } i \in \{1, 2, 3\},
\]
where $n'$ is a positive scaling factor, $V_\text{T}$ and $I_S$ are the thermal voltage and the saturation current of a BJT respectively, and $C_S$ ($i \in \{1, 2, 3\}$) are positive constants, is not a diffeomorphism, and may be applied to (2) in a formal sense only. This yields:
\[
\begin{align*}
C_S \dot{v}_{C_{S1}} &= -I_o \exp \left( \frac{v_{C_{S3}} - v_{C_{S1}}}{V_\text{T}} \right) + i' \\
C_S \dot{v}_{C_{S2}} &= I_o \exp \left( \frac{v_{C_{S1}} - v_{C_{S2}}}{V_\text{T}} \right) - I_o \exp \left( \frac{v_{C_{S3}} - v_{C_{S2}}}{V_\text{T}} \right) - aI_o \\
C_S \dot{v}_{C_{S3}} &= I_o \exp \left( \frac{v_{C_{S1}} - v_{C_{S3}}}{V_\text{T}} \right) - aI_o
\end{align*}
\]
where
\[
i' = \frac{T'}{nT_S} \exp \left( \frac{v_{C_{S1}}}{V_\text{T}} \right) \left( 2 \tanh \left( \frac{nT'_3}{I_o} \exp \left( \frac{v_{C_{S3}}}{V_\text{T}} \right) - m' \right) + m' \right),
\]
and, assuming $C_S = C_{S2} = C_{S3} = C_1$, $I_o = C_3 V_\text{CT} \omega_i$. The original system (2) and the mapped one (4) are not dynamically equivalent. Nonetheless, using the large-signal static forward-active mode exponential BJT relationship between the collector current $i_C$ and the base-emitter voltage $v_{be}$, i.e.
\[
i_i = I_S \exp(\pm v_{be} V_\text{T}^{-1}),
\]
where the positive (negative) sign corresponds to a npn (pnp) BJT, a log-domain oscillator may be still devised as a formal transistor-level realization of (4). In fact, the terms on the right-hand side of the $i$-th equation in (4) ($i \in \{1, 2, 3\}$) may be interpreted as currents flowing through a grounded capacitor of value $C_S$ and voltage $v_{C_{Si}}$ across it. All these terms but two may be formally implemented using either the positive or the negative log-domain only-npn cell of Figure 1, where $V_{CC}$ is the supply voltage, $i_i$ and $v_i$ ($v_C$ and $v_o$) denote the input and output currents (voltages) respectively, and $I_o$ and $I_C$ are constant current sources. In fact, application of basic principles from circuit theory to each cell in Figure 1 yields:
\[
i_i = I_o \exp(\pm (v_i - v_o) V_\text{T}^{-1}),
\]
where the positive (negative) sign corresponds to the positive (negative) cell.

It can be shown that the emitter current of BJT $Q_{14}$ (i.e. $I_{e14}$) in the circuit of Figure 2, in which four transistor loop $Q_{14}$-$Q_{10}$ and biasing current source $I_e$ formally implement the tanh nonlinearity, equals $i'$ if $I_o = m'I_e/n'$, $R = 2V_T n'/I_o$, $I_2 = 2I_o/n'$, and $I_3 = (-1 + m'/n') I_o$.

Finally, a grounded constant current source, i.e. $I_o$, formally realizes the constant term $aI_o$. The complete physically realizable log-domain oscillator from [4] is shown in Figure 3. Surprisingly, its dynamics differ from those of (4), and were explored in detail in [4]. However, a valid analytical model of the chaotic behaviour of the log-domain oscillator, not provided in [4], is necessary. The reason behind the inability of the BJT static model, upon which (4) is based, to track the dynamics of the oscillator of Figure 3 is due to the unexpectedly strong impact of the tiny BJT parasitic capacitances on the behaviour under study. This is not uncommon in log-domain circuits [7].

### 2.2. Modelling the Circuit Dynamics

#### 2.2.1 Dynamic BJT Model and Characteristic Equations of the Sub-circuits

Each transistor in the circuit of Figure 3 is first replaced with its Ebers-Moll large-signal static model. Then a parasitic capacitance of value $C_P$ is added in parallel to each diode in the model, as shown in Figures 4(a) and 4(b) for a pnp and a npn BJT respectively. Because of the complexity of the BJT model of Figure 4, there is a need to break down the difficult problem of modeling the large
respectively. The pseudo-state vector $i_j$ employs two P and two N sub-circuits. The input (domain cells of Figures 1(a)-(b). The oscillator of Figure 3 labels P and N denote the positive and negative log-domain cells (9) for $j=6$ and 0 otherwise, and $l_{o,j}=[0,1]c_{o,j-1}$ for $j=6$ and 0 otherwise. Application of Kirchhoff’s Voltage Law to the $j$-th sub-circuit reveals that the pseudo-state variables of the corresponding dynamical system are constrained to lie at all times on the affine invariant manifold described by:

$$D_j^T y_j = d_{o,j},$$

where $d_{o,j} \in \mathbb{R}^n$. $p_j$ is 0 for $j \in \{1, 2, 6\}$, 2 for $j \in \{3, 4, 8\}, 3$ for $j \in \{5, 7\}$, and $D_j^T$ is a suitable matrix.

### 2.2.2 Modelling the Interconnections and Final Equations

The pseudo-state vector of the dynamical system corresponding to the complete oscillator is $y'=[y_1', ..., y_n']^T \in \mathbb{R}^n$, where $M=n_1+ ... + n_n=43$. Noting that $k_1+ ... + k_n=2n-1=15$, defining the input and output vectors of the overall system, i.e. $i$ and $v$, as $i=[i_1', ..., i_n']^T \in \mathbb{R}^{2n-1}$ and $v=[v_1', ..., v_n']^T \in \mathbb{R}^{2n-1}$, and letting $F(y')=[F(y_1')^T, ..., F(y_n')^T]^T$, the collection of the characteristic equations of the $n$ sub-circuits yields:

$$\frac{d y'}{d \tau} = AF(y') + b_o + Bi,$$

where $A$, $b_o$, $B$, and $c_o$ are suitable matrices. Recalling (8), the affine invariant manifold on which solutions to (9) must lie at all times is given by:

$$D_j^T y_j = d_{o,j},$$

where $q=p_1+ ... + p_n=12$, $d_o \in \mathbb{R}^n$, and $D_j^T$ is a suitable matrix. Equations describing the interconnections among the $m$ grounded capacitors, current source $I_e$, resistor $R$, and the $n$ sub-circuits are here derived. Equating voltages at nodes $\oplus$, $\ominus$, $\ominus$, $\ominus$, and $\ominus$ in Figure 3 yields $2n-2m-1$ linear equations in $v$:

$$K_v = 0,$$

where $K$ is a suitable matrix. A further $m$ linear equations in $v$ and $i$ are established by the defining equations of the grounded capacitors:

$$\frac{H}{C_p} \frac{dv}{d\tau} + Li = \rho L_i,$$

for suitable matrices $H$, $L$, and $\rho$. Finally, applying Kirchhoff’s Current Law to nodes $\oplus$, $\ominus$, and $\ominus$ in Figure 3 a further $m$ linear equations in $v$ and $i$ are obtained:

$$\Lambda v + \Pi i = 0,$$

where $\Lambda$ and $\Pi$ are suitable matrices. Solving (11)-(13) for $i$, and inserting the solution into (9) yields:

$$\frac{d y'}{d \tau} = M_1 y' + M_2 F(y') + M_o.$$
where $M_1, M_2$, and $M_o$ are suitable matrices. From (9) and (11) it follows that solutions to (14) must satisfy at all times the following $2n-2m-1$ constraints

$$KC^T y' = -Kc_o,$$

(15)

in addition to the $q$ constraints imposed by (10). Therefore, defining $s = 2n-2m-1+q = 21$, the $M$-dimensional system (14) is constrained to operate at all times of an affine invariant manifold of dimension $M-s = 22$. Forming a vector $z$ with $M-s$ linearly independent components of $y'$, and using (10) and (15) to express the remaining $s$ components of $y'$ as a linear combination of the elements of $z$, (14) takes the final non-redundant form:

$$\frac{dz}{dt} = T_1 z + T_2 F(z) + T_3 (W_i z + W_o) + T_o,$$

(16)

where $T_1, T_2, T_3, W_i, W_o,$ and $T_o$ are suitable matrices.

3 Model Validation

In this section the BJT dynamic model of Figure 4 is validated by showing the dynamical equivalence between a numerical solutions to Equation (16) and the PSpice simulations of the oscillator of Figure 3 for a special set of circuit parameters yielding chaotic behaviour. In both PSpice and Matlab simulation environments the circuit parameters are set to: $V_{CC} = 3$ V, $I_o = 100$ $\mu$A, $I_h = 500$ $\mu$A, $C_3 = 0.3$ nF, $m' = 5$, $n' = 2$, and, from definitions given in 2.1, $I_i = 250$ $\mu$A, $R = 1$ K$\Omega$, (taking $V_T = 25$ mV), $I_2 = 100$ $\mu$A, and $I_3 = 150$ $\mu$A. The model parameters of npn (pnp) BJTs are chosen as those of practical transistor 2N3904 (2N3905). In the numerical integration of Equation (16) each BJT parasitic capacitance is set to $C_p = 10$ pF. Figure 5 depicts the chaotic attractor of the dynamical system, observed in PSpice for $a = 0.2$ and in Matlab for $a = 0.19$. Figure 6 shows the dynamic nonlinearity introduced by negative log-domain cell $N_2$ in Figure 3. It is clear that $N_2$ does not act as a static exponential nonlinearity. Therefore (6) does not model its input-output behaviour. This explains the inability of mapped system (4), based on the BJT static model described by (5), to capture the dynamics of the oscillator.

![Figure 5](image)

Figure 5. Chaotic attractor: (a) circuit simulation for $a = 0.2$; (b) numerical integration of (16) for $a = 0.19$.

![Figure 6](image)

Figure 6. Dynamic nonlinearity introduced by subcircuit $N_2$: (a) circuit simulation for $a = 0.2$; (b) numerical integration of (16) for $a = 0.19$.

4. Conclusions

Contrary to common expectations, due to their intrinsic internal nonlinearity, the dynamics of log-domain circuits may sometimes be strongly affected by the tiny BJT parasitic capacitances. This is also the reason behind the failure of the transistor large-signal static model to explain the chaotic behaviour of a recently presented log-domain oscillator. The simplest BJT large-signal dynamic model was used here to derive a set of nonlinear differential equations capturing the chaotic dynamics of the oscillator. The BJT model was then validated by comparing circuit simulations with numerical integrations of the equations.

References


Packet Traffic Analysis in a Complex Network with Multiple Power-Law Properties

Yongxiang Xia, Chi K. Tse and Francis C.M. Lau

Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong
http://chaos.eie.polyu.edu.hk

Abstract—Traffic dynamics is an important consideration in the design and planning of communication networks. In this paper we analyze the traffic congestions in a complex communication network. The scale-free (SF) network is used as the network model. Furthermore, we use power-law distributions to characterize the packet generation probabilities, packet delivery rates and buffer sizes in different nodes. Traffic congestion of this complex traffic model is simulated and the effects of network structure and traffic parameters on the traffic dynamics are studied.

1. Introduction

Traffic analysis of communication networks is one of the key issues of communication engineering [1]. Because of the large scale and complicated structure of communication networks, a well defined model, which can characterize the main properties of network structure, is necessary for the study of network traffic.

A widely used network model is the random graph. In the model developed by Erdos and Renyi [2], a random graph is built by connecting each pair of nodes with equal probability. This model is simple and clearly shows the stochastic characteristics of networks. Also, networks based on random graphs are homogeneous, meaning that the contribution of each node is statistically identical.

With the advent of computer technology, many empirical data from real-world networks have been collected. These data indicate that topologies of many real networks are at significant variance from the random graph model. Furthermore, recent research has shown that many real networks, including many communication networks, are typically heterogeneous networks having a power-law degree distribution:

\[ P(k_i) \sim k_i^{-\gamma} \]  

(1)

where \( k_i \) is the degree (i.e., the number of connections) of node \( i \) and \( \gamma \) is the characteristic exponent [3, 4]. Such networks are called scale-free networks. Equation (1) shows that in the network, there are a small number of hubs, each having a large number of connections. At the same time, there are a large number of nodes, each of which only has a few connections.

The discovery of SF property in real world networks has spawned the study of the structure and dynamics of complex networks [5]. In particular, the traffic dynamics of SF networks has received a great deal of attention [6, 7, 8, 9, 10, 11, 12]. Early research was initiated by statistical physicists who first observed that many real communication networks are actually SF networks. Some basic traffic problems were also studied, including the cascade-based failure, phase transition of traffic dynamics, and the avalanche dynamics of traffic congestion, etc. Most research in this phase assumes some ideal preconditions. For example, the queue length (or the buffer size) at each node has been assumed to be infinitely long [10]. However, in any real communication network, the buffer size cannot be infinite. Recently, as the research progresses, some traffic models with more realistic considerations are considered [11, 12].

In our previous work [11], we have shown the effect of network structure on traffic congestions. In order to emphasize the importance of hubs, we assumed in [11] that the packet generation probability and the delivery capability for each node are both proportional to its degree. In this way, if a node has more connections, it can generate more packets and transmit more packets to their destinations.

In this paper, we will extend that preliminary model. First, the packet generation probability and the delivery capability are scaled with, but not necessary linearly proportional to, the degree. Second, we set the buffer size to scale with the degree. In this way, we get a more general model. The preliminary model proposed in [11] can be considered as a special case of the new model. Using this new model, we will analyze the traffic congestion with different parameter settings.

2. Network Structures

Our main aim is to study the traffic congestion in SF networks. Here we use the extended BA model to construct the SF network [13]. For comparison, we also consider the random graph because it is a widely used homogeneous network model. The performance comparison between them can help us understand the important role that the network structure plays in the flow of network traffic.

2.1. Random Graphs

The construction of random graphs is quite simple. Consider a network with \( N \) nodes. Each pair of nodes are con-
nected with probability $p$. Consequently, the total number of connections $n$ in the network is a random variable with a mean value equal to $E(n) = pN(N - 1)/2$. For each node, its degree $k$ is also a random variable. If $N$ is large enough, the degree distribution follows a Poisson distribution [3], i.e.,

$$P(k) = C_p^{N-1} p^k (1 - p)^{N-1-k} = \frac{(k-p)^{e-(k)}}{k!}$$  \hfill (2)

where $(k) \approx p(N - 1) \approx pN$ is the mean value of $k$. Since each pair of nodes are connected with equal probability, the random graph is a homogeneous network.

2.2. Scale-Free Networks

We use the evolution algorithm to construct the SF network [13]. The network starts with $m_0$ isolated nodes. At each time step, one of the following three processes takes place with a certain probability.

1. With probability $p$, $m$ ($m < m_0$) new connections are added into the network. To do this, for each new connection we randomly choose a node as one end. The node $i$ is selected as the other end of the connection with a probability of

$$\prod(k_i) = \frac{k_i + 1}{\sum_j (k_j + 1)}.$$  \hfill (3)

2. With probability $q$, $m$ existing connections are rewired. To do this, we randomly choose a node (say node $i$) and a connection from this node (say the connection from node $i$ to node $j$). We remove this connection and rewire it between node $i$ to $j'$, where $j'$ is randomly chosen with probability $\prod(k_j)$ given by (3).

3. With probability $1 - p - q$, a new node with $m$ new connections is added into the network. The $m$ new connections are connected to existing node $j$ with probability $\prod(k_j)$ given by (3).

This algorithm simulates the evolution of real networks such as the Internet or airline networks. Consider the airline network as an example. At the beginning, there are only a few airports in the network. Some new flight routes are set up between existing airports. With the development of air-travel market, new airports are built and connected to existing airports with new routes; more routes are also added between existing airports; some existing routes may be cancelled and reconnected to other airports. All these processes contribute to today’s air traffic network.

Based on this construction method, the degree distribution of the evolved network has a generalized power-law form [13]:

$$P(k) \propto [k + \kappa(p, q, m)]^{-\gamma(p, q, m)}$$  \hfill (4)

where $\kappa(p, q, m) = 1 + (p - q)\left(\frac{2m(1-q)}{1-p-q} + 1\right)$ and $\gamma(p, q, m) = 1 + \frac{2m(1-q)}{m - 1-p-q}$. The degree distribution of a SF network has a power-law tail. In this network, a small number of nodes have a very large number of connections. It indicates that the effects of different nodes in the SF networks are not identical. In other words, the SF network is a heterogeneous network.

3. Traffic algorithm

The algorithm for constructing the traffic model in this paper is similar to the routing algorithm in computer networks, which can be summarized as follows:

1. Each node in the network has a buffer. The buffer size for node $i$ is $B_i$.

2. Packet generation: At each time step, new packets are generated. The probability for node $i$ to generate a packet is $\lambda_i$. When a packet is generated, its destination is chosen randomly. Then, the shortest path from the source to the destination is determined. If there are more than one shortest path, we randomly choose one. The newly generated packet is placed at the end of the buffer of this node. If the buffer is already full, the newly generated packet is dropped.

3. Packet transmission: At each time step, node $i$ delivers the first $\mu_i$ packets in its buffer one step towards their destinations. For each packet, if the next node is not its destination, the buffer status of that node is checked. If the buffer is full, the packet is dropped. Otherwise, the packet is added to the end of the buffer.

4. Packet removal: A packet is removed from the system when it reaches its destination.

In this traffic model, a packet drop indicates that a congestion occurs. If we define the packet drop probability $P_d$ as

$$P_d = \frac{\text{average number of dropped packets}}{\text{average number of generated packets}}$$  \hfill (5)

the value of $P_d$ can be used to evaluate the system performance. A higher drop probability indicates that a larger percentage of packets cannot reach their destinations. As a result, the quality of service of the communication network becomes worse.

4. Parameter Settings

In the traffic model, the packet generation rates for different nodes are different. Generally speaking, the more popular a node is, the more packets are generated from this node. In our model, the degree of a node can be considered as an index of popularity of that node. Thus, we assume that the packet generation probability for node $i$ is

$$\lambda_i = \alpha(k_i)^{\gamma}, \lambda_i \leq 1,$$  \hfill (6)

where $\alpha$ is a proportionality constant. In this way, the packet generation probability also follows a power law.
Table 1: Values of $\lambda_i$ for different $k_i$ and $\gamma_1$. $\gamma_1$ increases from 0 to 1.5. A scale-free network is considered, where $N = 1000$ and $\gamma \approx 2.4$.

<table>
<thead>
<tr>
<th>$k_i$</th>
<th>$\gamma_1 = 0$</th>
<th>$\gamma_1 = 0.1$</th>
<th>$\gamma_1 = 0.5$</th>
<th>$\gamma_1 = 1$</th>
<th>$\gamma_1 = 1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>0.100</td>
<td>0.132</td>
<td>0.362</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0.100</td>
<td>0.118</td>
<td>0.209</td>
<td>0.350</td>
<td>0.443</td>
</tr>
<tr>
<td>1</td>
<td>0.100</td>
<td>0.094</td>
<td>0.066</td>
<td>0.035</td>
<td>0.014</td>
</tr>
</tbody>
</table>

The value of exponent $\gamma_1$ is very important in the packet generation process. When the network structure is fixed, it determines the distribution of packet generation rate. In particular, when $\gamma_1 = 0$, $\lambda_i$ ($i = 1, 2, \cdots, N$) are identical, which means that all nodes generate packets with the same rate. When $\gamma_1 > 0$, the packet generation probability follows a power-law distribution. A hub (i.e., a node with a large number of connections) generates packets at a much larger rate than most of the other nodes. Specifically, when $\gamma_1 = 1$, (6) becomes

$$\lambda_i = \alpha k_i,$$

which corresponds to the case where the packet generation probability is linearly proportional to the degree. Table 1 gives an example of how the $\lambda_i$ distribution changes with $\gamma_1$. We consider a SF network with 1000 nodes and $\gamma = 2.4$. The conditions of three nodes are studied: a hub node with 30 connections, a non-hub node with only 1 connection, and a node with 10 connections, just between the above two. When $\gamma_1 = 0$, the packet generation probabilities for all nodes are 0.100. When $\gamma_1 > 0$, there is a gap between the packet generation probabilities of the hub node and the non-hub node. This gap grows with $\gamma_1$. For $\gamma_1 = 0.1$, it is only 0.038, whereas for $\gamma_1 = 1.5$, it becomes 0.986.

Similarly, we assume that the packet delivery rate of a node is also scaled with its degree, i.e.,

$$\mu_i = [1 + \beta (k_i)^{\gamma}],$\quad (8)$$

where $\beta$ is a constant. The exponent $\gamma_2$ determines the distribution of packet delivery capabilities among different nodes. When $\gamma_2 > 0$, the distribution of $\mu_i$ is power-law distributed, which means that nodes with larger degrees have higher delivery capabilities.

The buffer size of node $i$ is given by

$$B_i = [1 + \delta (k_i)^{\gamma}],'\quad (9)$$

where $\delta$ is a constant. The exponent $\gamma_3$ determines the distribution of buffer sizes among different nodes. When $\gamma_3 > 0$, the distribution of $B_i$ is power-law distributed, which means that nodes with larger degrees have larger buffer sizes.

The above settings emphasize the important role of hubs. These nodes are more powerful than others. Here, by “more powerful” we mean that the node can generate and deliver more packets and has a larger buffer size.

5. Simulations

The simulation parameters are set as follows:

- $N = 1000$ and $\gamma \approx 2.4$ used in the network construction;
- $\sum_{i=1}^{N} \alpha_i \approx 100, \sum_{i=1}^{N} \mu_i \approx 2100$ and $\sum_{i=1}^{N} B_i \approx 10000$ used in the traffic model.

The approximations are consequences of the random nature of the complex network construction and the traffic dynamics. For the SF network, $(k) \approx 2.9$ when $\gamma \approx 2.4$. In order to show the effect of each of the characteristic exponents $\gamma_1, \gamma_2$ and $\gamma_3$ separately, for each set of simulations, we alter the value of only one exponent while fixing the other two exponents at 1. When the value of a characteristic exponent varies, the corresponding proportionality constant ($\alpha, \beta$ or $\delta$) will change accordingly such that the conditions stated above are satisfied.

Figure 1 plots the drop probability versus $\gamma_1$. By varying $\gamma_1$, the relative distribution of packet generation probability among the nodes is changed. The simulation result indicates that such a variation does not affect the network performance.

*Observation 1*: Regardless of whether the packets are more likely to be generated by hub nodes or by non-hub nodes, the resulting drop probability is very similar.

Figure 2 shows the effect of $\gamma_2$ on the drop probability. As $\gamma_2$ increases, the drop probability first decreases and then stays at a small value.

*Observation 2*: The above result suggests that assigning more delivery capability to hubs benefits the traffic performance.

The effect of $\gamma_3$ on the drop probability is shown in Fig. 3. For the SF network, as $\gamma_3$ increases, more and more buffers are assigned to nodes with larger degrees. These nodes are hubs and a large number of packets go through them. Thus, increasing the buffer sizes of these nodes can decrease the drop probability effectively. However, if the
hubs have acquired enough buffers, further increasing their buffer sizes has little effect on the traffic performance. If we keep increasing γ₃, buffers keep accumulating to nodes with large degrees. At the same time, buffer sizes of nodes with small degrees may be inadequate, making the drop probability increase. For the random graph, the drop probability always remains at a very low level.

**Observation 3:** The above simulation result suggests that assigning adequate buffer resource to hubs benefits the traffic performance. But assigning too many buffers to hubs may waste system resource and degrade the performance. Hence, an optimal point has to be found for effective network management.

In all the simulations, the drop probability for the SF network is always much higher than that of the random graph. This is because the SF network is heterogeneous, in which hubs become the traffic bottleneck. In comparison, the traffic in a random graph is distributed more uniformly, causing fewer packet drops.

**Observation 4:** In real-world networks which are scale-free, the drop probability is much higher than that of the random-graph model.

6. Conclusions

In this paper we study the traffic flow in a scale-free network. In the traffic model, the packet generation probability, the packet delivery capability, and the buffer size are all power-law distributed. We study the effect of these three power-law distributions on the traffic congestion performance. The simulation results show that assigning higher packet delivery capability or reasonably larger buffer size to those nodes which have larger degrees can effectively enhance the traffic performance. Also, random-graph models consistently underestimate the traffic volume of real-world networks.

Acknowledgment

This work was supported by the Hong Kong Polytechnic University under projects G-YE38 and 1-BBZA.

References


State-feedback control of complex dynamical networks

Zhengping Fan*a, Francesco Sorrentino*b Mario di Bernardo*b and Guanrong Chen*a

* Department of Electronic Engineering, City University of Hong Kong, China
* Department of Systems and Computer Science, University of Naples Federico II, Napoli, 80125, Italy

Abstract– The control issue of complex dynamical networks is considered in this paper, in the case where a control action is locally applied to all the network dynamical nodes. By making use of classical state-feedback control techniques, a complex dynamical network is stabilized to one of its equilibrium points. Both theoretical analysis and numerical simulations are reported in this paper.

1. Introduction

Complex dynamical networks are composed of a large number of dynamical nodes, which may be coupled to each other through diffusion, convection, conduction, etc. [1]. Typical examples include cellular neural networks [2], coupled map lattices [3], arrays of Josephon junctions [4], and coupled chemical reactors [5]. Obviously, the “global” behaviors of such complex networks are determined by both the dynamics at every node and the interactions among them (for instance, electrical activity in neural or cardiac cells is the result of a large number of ionic currents and signaling events [6]).

The complicated dynamical behaviors of complex dynamical networks may result harmful in many cases, therefore controlling the dynamics taking place over complex networks is important. As a matter of fact, the issue of controlling complex dynamical networks has become a focal topic in the control community in the past decade. However, most of the existing works on controlling complex dynamical networks have been concentrated on networks with completely regular topological structures, such as chains, grids, lattices, and fully-connected graphs. The main benefit in using these simple architectures is that it allows for focusing on the complexity caused by the nonlinear dynamics at the nodes without considering additional complexity in the network structure itself. The effects of the network topology on its dynamics, however, cannot be ignored in many cases, as well as its rapid structural evolution over time. For example, it was found [7] that the multicast resource reservation styles in the Internet are quite different from those with the simple linear, tree, or star-shaped topologies. It was also reported [8] that the synchronizability of scale-free networks, characterized by a power-law degree distribution, is strongly affected by the exponent of the power-law.

In this paper, the issue of controlling complex dynamical networks is considered. Firstly, a linearly coupled dynamical network is introduced. Then, based on the conventional state-feedback control method, a sufficient condition on controlling a complex dynamical network to its equilibrium point is obtained. Finally, numerical simulations are given to verify the effectiveness of the given method.

2. Problem description

Suppose that a complex network consists of \( N \) linearly and diffusively coupled identical nodes, with each node being an \( n \)-dimensional dynamical system. The state equations of this dynamical network are given by

\[
\dot{x}_i = f(x_i) + c \sum_{j=1}^{n} a_{ij} \Gamma x_j, \quad i = 1, 2, \cdots, N. \tag{1}
\]

Here, \( x_i = (x_{i1}, x_{i2}, \ldots, x_{in})^T \in \mathbb{R}^n \) are the state variables of node \( i \), the constant \( c > 0 \) represents the coupling strength. \( \Gamma = (\gamma_{ij}) \in \mathbb{R}^{n \times n} \) is a symmetric matrix linking coupled variables, and if for some pairs \( (i,j) \), \( 1 \leq i,j \leq n \), \( \gamma_{ij} \neq 0 \), it means two coupled nodes are linked through their \( i \)-th and \( j \)-th state variables, respectively. Here, we assume that each pair of coupled oscillators is linked through the complete set of identical sub-state variables, i.e., \( \Gamma = \text{diag}(1, \ldots, 1) \).

In network (1), the symmetric coupling matrix \( A = (a_{ij}) \in \mathbb{R}^{N \times N} \) represents the connectivity configuration of an undirected complex network. If there is a connection between node \( i \) and node \( j \) (\( i \neq j \)), then \( a_{ij} = a_{ji} = 1 \); otherwise, \( a_{ij} = a_{ji} = 0 \) (\( i \neq j \)).

The degree \( k_i \) of node \( i \) is defined to be the number of connections incident in it:

\[
\sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_{ji} = k_i, \quad i = 1, 2, \cdots, N. \tag{2}
\]

Let the diagonal elements be \( a_{ii} = -k_i, i = 1, 2, \cdots, N \). Suppose the network is connected in the sense of having
no isolated clusters, i.e., the symmetric matrix $A$ is irreducible. We know that zero is an eigenvalue of $A$ with multiplicity 1, and other eigenvalues of $A$ are strictly negative in the order: $0 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_N$.

The objective here is to stabilize network (1) onto an equilibrium point of the network, in the sense that $x_i(t) = x_{i+1}(t) = \cdots = x_N(t) \rightarrow \bar{x}$, as $t \rightarrow \infty$, (3) where $\bar{x} \in \mathbb{R}^n$ is a solution of an isolated node, satisfying $\dot{\bar{x}} = f(\bar{x})$.

Without loss of generality, we assume that network (1) satisfies the following assumption:

**Assumption:** For any isolated node, there exists a non-negative constant $\beta$ such that

$$(x - y)^T (f(x) - f(y)) \leq \beta (x - y)^T (x - y), \quad \text{where} \quad \beta > 0 \quad \text{is a constant.}$$

### 3. Main result

To achieve the goal (3), the state feedback control strategy is applied to stabilize network (1). Therefore, the controlled network (1) can be described by

$$\dot{x}_i = f(x_i) + c \sum_{j=1}^{N} a_{ij} x_j + u_i, \quad i = 1, 2, \cdots, N. \quad (5)$$

Here $u_i$ is the controller of node $i$, described by

$$u_i = -k(x_i - \bar{x}), \quad i = 1, 2, \cdots, N.$$ (6)

where $k > 0$ is the constant control strength.

Define $e_i(t) = x_i(t) - \bar{x}$; then one can have

$$\dot{e}_i = f(x_i) - f(\bar{x}) + c \sum_{j=1}^{N} a_{ij} e_j - k e_i, \quad i = 1, 2, \cdots, N. \quad (7)$$

If the given assumption holds, we can get a sufficient condition for stabilizing network (1):

**Theorem:** Assume the given assumption holds. Network (5) is globally stabilized to $\bar{x}$ if

$$\lambda_{\max}((\beta - k)I + cA) < 0,$$ (8)

where $\lambda_{\max}(M)$ is the highest eigenvalue of matrix $M$.

Proof: Construct a Lyapunov functional of the form

$$V = \frac{1}{2} \sum_{i=1}^{N} e_i^T e_i \quad (9)$$

The time derivative of (9) along trajectories of (7) is

$$\dot{V} = \sum_{i=1}^{N} e_i^T (f(x_i) - f(\bar{x})) + \sum_{i=1}^{N} e_i^T \left( c \sum_{j=1}^{N} a_{ij} x_j \right) - k \sum_{i=1}^{N} e_i^T e_i.$$

Define

$$\bar{e}_j = \begin{bmatrix} e_{1j} \\ e_{2j} \\ \vdots \\ e_{Nj} \end{bmatrix}, \quad j = 1, 2, \cdots, N,$$

one has

$$\dot{V} \leq \sum_{i=1}^{N} \left[ (\beta - k) e_i^T e_i + c \sum_{j=1}^{N} e_i^T a_{ij} e_j \right] - k \sum_{i=1}^{N} e_i^T e_i \leq \beta \sum_{i=1}^{N} \left[ e_i^T e_i \right] + c \sum_{j=1}^{N} \left[ e_j^T A e_j \right] - k \sum_{i=1}^{N} e_i^T e_i$$

$$= \sum_{j=1}^{N} e_j^T \left( (\beta - k)I + cA \right) e_j.$$

Obviously, $\dot{V} < 0$ if

$$\lambda_{\max}((\beta - k)I + cA) < 0.$$ (10)

This completes the proof of the theorem.

**Corollary:** Assume the given assumption holds.

Network (5) is globally stabilized to $\bar{x}$ if $k \geq \beta.$ (11)

Proof: Define $\bar{e}_j = P v_j \quad (j = 1, 2, \cdots, n)$ where matrix $P$ is such that $P^T AP = \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)$, one has

$$\bar{e}_j^T \left( (\beta - k)I + cA \right) e_j$$

$$= v_j^T ((\beta - k)I + c \times \text{diag}(\lambda_1, \lambda_2, \cdots, \lambda_N)) v_j$$

$$= \sum_{j=1}^{N} (\beta - k + c \lambda_j) v_j^2 \quad (j = 1, 2, \cdots, n).$$

Clearly, $\dot{V} < 0$ if $k \geq \beta.$

This completes the proof of the corollary.

### 4. Numerical Simulations

Consider a chaotic dynamical network of Lorenz systems, where a single Lorenz oscillator is described by

$$\dot{x}_i = \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = f(x) = \begin{bmatrix} p_1 (x_2 - x_1) \\ p_2 x_1 - x_1 x_3 - x_2 \\ x_1 x_2 - p_3 x_3 \end{bmatrix}, \quad (12)$$

with the parameter set of $p_1 = 10, \ p_2 = \frac{8}{3}, \ p_3 = 28$, and $\beta = 60$ in the assumption.

We generate a scale-free network with the BA model [9], which has 200 nodes with parameters $m = m_0 = 2$, and set the coupling strength $c = 0.01$. To stabilize such a BA network to the homogenous state $\bar{x} = (0,0,0)$, we set the control strength $k = 62.5$. Since
\[ \lambda_{\text{max}} \left( (\beta - k)I + cA \right) = -2.5 < 0, \] 
the condition in the theorem is satisfied.

States \( x \) and \( y \) of the highest degree node in the network are shown in Figs. 1-2. From 0 to 40 seconds, we do not have any control to nodes. Consequently, chaotic phenomenon is observed. After 40 seconds, all nodes are controlled, and then they tend to \( x = (0,0,0) \) quickly.

Similarly, we construct a small-world network with the NW small-world model [10] with network parameters \( N = 200, K = 2 \) and \( p = 0.01 \). We set the coupling strength \( c = 0.01 \) and the control strength \( k = 63 \). From Figs. 3-4, one can see that the network is well stabilized to its equilibrium point \( x = (0,0,0) \).

5. Conclusion

The stabilization of complex dynamical networks onto a homogenous state has been studied in this paper. Based on a classical state-feedback control strategy, a sufficient condition was obtained to stabilize a dynamical network with a fixed coupling strength. Numerical simulations have shown that the proposed method can be used to stabilize such networks as BA scale-free and NW small-world dynamical networks. However, a real network is usually very large, which makes almost impossible to control all the nodes of the network. Therefore, the pinning control strategy [11-12], in which only a portion of nodes in the network is controlled, may benefit the control of many real-world dynamical networks. Another important aspect is represented by the role played by the network topology on the effectiveness of the control action (in an analogous way to the propensity of complex networks to being synchronized, depending on some of their structural properties [13-14]). These topics will be discussed in our future research.
Acknowledgements

This research was supported by the Hong Kong Research Grants Council under the CERG Grant 1114/05E.

References

Growing Self-Organizing Maps for Network Construction

Toshimichi Saito†, Takeshi Ehara†, Satoshi Suzuki† and Hiroyuki Torikai†

†EECE Dept, HOSEI University, Koganei-shi, Tokyo, 184-8584, Japan, tsaito@k.hosei.ac.jp

Abstract—This paper discusses application of self-organizing maps to network construction. The maps have growing cell structure for flexible adaptation to environment and have applied to two examples. The first one is traveling sales person problem, a typical combinatorial optimization problem. The map has ring topology and city positions are applied successively as input data. Although the algorithm is very simple we can find almost optimum solutions. The second one is construction of complex network. Nodes are applied successively as input data and the algorithm can adapt to the dynamic input space flexibly. Performing basic numerical experiments we show that our algorithm can generate small-world networks.

1. Introduction

This paper discusses an approach to fuse complex networks and self-organizing maps (ab. SOMs ). There has been increasing interest in the study of complex networks, including small-world and scale-free networks [1] - [3]. The study relates to both fundamental and application objects including construction of complex networks, modeling physical networks, synchronization in networks, communications, environment and signal separation. On the other hand, SOMs have been studied intensively as typical unsupervised learning systems [4]. In order to enhance adaptive function, SOMs are improved to use growing cells structure ( ab. GCS ) [5]: the maps can change size and topology depending on features of input data. The GCS has a variety of applications including data visualization, pattern classification, vector quantization and image skeletonization [6] - [10]. If GCS has 1D topology, the cell structure is a kind of graph and the GCS relates deeply to construction of networks.

This paper discusses application of the GCS to network construction based on two examples. The first one is the traveling sales person problem ( ab. TSP ) [11]-[14]. The TSP is well known as a typical combinatorial optimization problem and the task is to find the shortest possible tour through the set of cities that passes through each city exactly once. For this problem, we use GCS having ring topology and city positions are applied successively as input data. Our algorithm is simple and can be controlled by only two parameters [15], [16]. Performing basic numerical experiments we show that our algorithm can find almost optimum tour.

The second example is construction of small-world networks. The concept of the small-world networks ( ab. SWN [11] ) has been introduced in order to interpolate between regular and random networks as depicted in Figure 1. In order to characterize networks two basic measures are known: the average path length $L$ and cluster coefficient $C$ [1][2]. Roughly speaking, $L$ measures typical separation between nodes and $C$ measures the cliquishness of a typical neighborhood. The SWNs are characterized by small $L$ as a random graph and relatively large $C$ as a regular lattice. Such networks can be an interesting network models and are probably generic for many natural networks including power grid of a nation, the collaboration graph of film actors and communication with fewer connections. Also, coupled dynamical systems having SWN-like topology display enhanced signal-propagation speed and ability of synchronization [1]-[3].

In order to study such network structure, some algorithm is necessary to generate a desired network for a set of nodes. In a typical algorithm by Watts and Strogatz [2] given regular lattices are reconnected following some probability without learning. Our algorithm is based on GCS and can learn adaptively to dynamic input space [17], [18]. Nodes for networks are applied successively as input data and the number of nodes can be changed depending on the learning history. Performing basic numerical experiments we show that our algorithm can generate small-world networks if control parameters are selected suitably. These results provide basic information to develop effective application of GCSs to constructing desired networks in communications system, traffic network and so on.

2. Traveling Salesperson Problem

The GCS consists of cells and the number of cells is time-variant. Let $i$ be a discrete time, $i$ be index of a cell and let $N(t)$ be the number of cells at time $t$. As illustrated in Figure 2 the GCS has ring topology and each cell is con-
connected to the both sides cells: \( i \in \{1, \cdots, N(t)\} \) and \( i \) is modulus \( N(t) \). In the figure \( X \) is a 2-dimensional input space and \( A \) is a discrete space of cells. The \( i \)-th cell is characterized by a 2-dimensional synaptic vector \( w_i(t) \in \mathbb{R}^2 \) and a signal counter \( C_i(t) \) for inspection of learning history. The data set \( D \) consists of location of cities and the object is finding the shortest tour. The algorithm consists of the following 7 steps.

**Step 1** (Initialization): Let \( t = 0, N(0) = 3 \) and let \( C_i(0) = 0 \). We then give initial value of \( w_i(0) \). For an example in Figure 3 \( w_i(0) \) corresponds to either apex of the triangle.

**Step 2** (Input signal): We select one point \( x \in D \) randomly and apply \( x \) as an input signal.

**Step 3** (Winner cell): Select a winner cell \( c \) whose synaptic vector \( w_c(t) \) is the closest to the input. If plural winners exist we select one of them randomly.

\[
\|x(t) - w_c(t)\| = \min_i \|x(t) - w_i(t)\|. 
\]  

**Step 4** (Update): The synaptic vectors of the winner and its neighbors are updated:

\[
w_i(t+1) = \begin{cases} 
    w_i(t) + \alpha(x(t) - w_i(t)) & \text{if } i \in \{c-1, c, c+1\} \\
    w_i(t) & \text{otherwise},
\end{cases}
\]  

(2)

The learning rate \( \alpha \) is the first control parameter of this algorithm. The signal counter of the winner is updated:

\[
C_i(t+1) = \begin{cases} 
   C_i(t) + 1 & \text{if } i = c \\
   C_i(t) & \text{otherwise}
\end{cases}
\]  

(3)

**Step 5** (Cell insertion): At every \( T_{\text{ins}} \) time we find the cell \( p \) having the maximum counter value.

\[
C_p(nT_{\text{ins}}) \geq C_i(nT_{\text{ins}}) \quad \text{for all } i.
\]

This is an inspection of a learning history. The insertion interval \( T_{\text{ins}} \) is the second control parameter of this algorithm. A novel cell \( q \) is inserted between \( p \) and its further neighbor \( r \). After the insertion, the number of cells increases:

\[
n(t+1) = N(t) + 1. \]

The synaptic vector of this new cell \( q \) is given by

\[
w_q(t+1) = 0.5(w_p(t) + w_r(t))
\]  

(4)

The counter values of cells \( p \) and \( q \) are given by

\[
C_p(t+1) = 0.5C_p(t), \quad C_q(t+1) = 0.5C_p(t).
\]  

(5)

**Step 6** (Finding a tour): If the number of cells exceeds the number of cities then we find the closest cells to each city. Every city can obtain the distinct closest cell hereby the tour route can be determined.

**Step 7** (Termination): Let \( t = t + 1 \). If \( t < T_{\text{max}} \) then go to step 2, otherwise the learning is terminated, where \( T_{\text{max}} \) is a learning limit.

Although this is a simple algorithm having only two control parameters \( \alpha \) and \( T_{\text{ins}} \), we can find good solution of TSP. Figure 3 recalls a basic example of 48 cities in [15]. As \( t \) increases the GCS grows and forms a tour route for \( t > 4800 \). The learning is terminated at \( t = 20000 \) when the GCS has 200 cells and forms a tour that is just 1.4 percent longer than the optimum route. It should be noted that we have improved this basic algorithm by applying parallel processing of plural GCSs [19], [20].

3. Network Construction

In this application, we need some different definitions from previous example. The GCS is represented by nodes \( n_i(t) \in \mathbb{R}^2 \) and edges \( e_{ij}(t) \) between nodes \( n_i(t) \) and \( n_j(t) \) where \( e_{ij}(t) \equiv 1 \) and \( e_{ij}(t) = 0 \) mean connection and disconnection between \( n_i(t) \) and \( n_j(t) \), respectively. A node \( n_i(t) \) has a signal counter \( C_i(t) \). Let \( n(t) \) be a set of nodes:

\[
n(t) \equiv \{n_1(t), \ldots, n_{N(t)}(t)\}.
\]

The \( i \)-th node \( n_i(t) \), \( i \in \{1, \ldots, N(t)\} \), is a 2-dimensional vector corresponding to a synaptic vector of GCS. Note that \( t, N(t) \) and \( C_i(t) \) are defined in Section 2. 

Figure 2: GCS having ring topology

Figure 3: TSP with 48 cities. \((N(0) = 3, \alpha = 0.1, T_{\text{ins}} = 100)\)
The data set $D$ is a set of points in some area: $D = \{d_1, \ldots, d_p\}$, $d_i \in \mathbb{R}^2$ where $p$ denotes the number of points. Note that $y_i$ corresponds to a position on the ground in some examples such as communication terminals and power grid. Our algorithm consists of the following 6 steps.

**Step 1** (Initialization): Let $t = 0$, $N(0) = 2$ and let $C_1(0) = C_2(0) = 0$. The two nodes $n_1(t)$ and $n_2(t)$ are connected ($e_1(t) = 1$) and located randomly.

**Step 2** (Input and two winners): We select a pair of points $(x_1(t), x_2(t))$ randomly from the set $D$ and apply it as input: $(x_1(t), x_2(t)) = D$, $x_i(t) \in \mathbb{R}^2$. We find the first winner node $n_{i1}(t)$ that is the closest to $x_1(t)$ and the second winner node $n_{i2}(t)$ that is the closest to $x_2(t)$.

$$
\|x_1(t) - n_{i1}(t)\| = \min_i \|x_1(t) - n_i(t)\|
$$

$$
\|x_2(t) - n_{i2}(t)\| = \min_i \|x_2(t) - n_i(t)\|
$$

(6)

If plural closest nodes exist to an input then we declare the node with smaller subscription $i$ as the winner. Note that the node with smaller subscription is older node introduced in Step 4.

**Step 3** (Update): The two winner nodes move to the inputs position and the signal counters $C_{i1}(t)$ and $C_{i2}(t)$ are updated.

$$
n_i(t+1) = \begin{cases} 
    x_i(t) & \text{if } i \in \{c_1, c_2\} \\
    n_i(t) & \text{otherwise}
\end{cases}
$$

(7)

$$
C_i(t+1) = \begin{cases} 
    C_i(t) + 1 & \text{if } i \in \{c_1, c_2\} \\
    C_i(t) & \text{otherwise}
\end{cases}
$$

(8)

**Step 4** (Cell generation): A node $n_i(t)$ is declared as a fire-node if its signal counter $C_i(t)$ reaches a threshold $th$. The threshold $th$ is the first parameter of this algorithm.

A new node $n_{N(t)+1}(t)$ is generated at the same position as the fire-node $n_i(t)$. The node $n_i(t)$ is connected to $n_{i1}(t)$ and its closest neighbor node $n_{ij}(t)$ as shown in Figure 4. The counter value of new node is shared with the fire-node.

$$
n_{N(t)+1}(t) = n_i(t) \text{ for } SC_i(t) = th
$$

$$
C_{N(t)+1}(t) = C_i(t) = 0.5th
$$

$$
e_{n_{N(t)+1}(t)}(t) = 1, \quad e_{n_{N(t)+1}(t)}(t) = 1
$$

The generation means $N(t+1) = N(t) + 1$. Note that there exists possibility of two fire-nodes and $N(t) = N(t)+2$.

**Step 5** (Cell deletion): If plural nodes exist over a lifetime $M$, a node with larger subscription (younger node) at the position is delated. The lifetime $M$ is the second parameter of this algorithm. The deletion means $N(t+1) = N(t) - 1$. Note that there exists possibility of two nodes delation $N(t+1) = N(t) - 2$.

**Step 6** (Termination): If all the points acquire nodes, $m(t) \cap D = D$, then learning is terminated. Otherwise let $t = t+1$ and go to Step 2.

In order to confirm the algorithm function we have performed a basic experiment for a data set of 50 points located on a circle as shown Figure 6. The input of a pair $(x_1(t), x_2(t))$ is selected randomly from the data set $D$ and are applied to the network. The two control parameters are fixed based on trials-and-errors:

$$
th = 10, \quad M = 3, \quad D = \{d_1, \ldots, d_{50}\}, \quad \|d_i\| = 1.
$$

Figure 6 shows the learning process. In the beginning there exist a small the number of nodes. They can move freely and long edge formation is possible. As the learning evolves, the number of nodes increases. The long edge

Figure 4: Node generation. a new node (mesh circle) is generated at the same position $g$ as the fire-node.

Figure 5: Node deletion: young node (mesh circle) at position $d$ is deleted.

Figure 6: Learning process for 50 points on a circle.
Table 1: Average path length $L$ and cluster coefficient $C$ for circular input space.

<table>
<thead>
<tr>
<th>#points</th>
<th>$L$ - GCS</th>
<th>$C$ - GCS</th>
<th>$L$ - WS</th>
<th>$C$ - WS</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>3.2</td>
<td>0.64</td>
<td>4.3</td>
<td>0.43</td>
</tr>
<tr>
<td>100</td>
<td>4.3</td>
<td>0.67</td>
<td>5.5</td>
<td>0.41</td>
</tr>
<tr>
<td>150</td>
<td>4.8</td>
<td>0.66</td>
<td>5.9</td>
<td>0.41</td>
</tr>
</tbody>
</table>

formation is to be hard and short edges are to be generated frequently. After the learning, the resulting network is evaluated using two basic measures: the average path length $L$ and cluster coefficient $C$. We have calculated $L$ and $C$ for data sets of 50, 100 and 150 points on the circle. Averaged values for 30 trials each are shown in Tables 1 [18]. The table also shows results by Watt-Strogatz method (ab. WS [2]) with reconnection probability $P = 0.07$. In the table, we can see that our algorithm constructs small-world-like networks characterized by smaller $L$ and larger $C$. Adjusting parameter values and initial conditions, our algorithm can construct a variety of complex networks adaptively to dynamic input data.

4. Conclusions

We have introduced simple self-organizing maps having growing cell structure and its application to construction of networks through two examples. In the first example we shows a simple algorithm that can find almost optimal tour of TSP. In the second example we show a flexible algorithm that can construct small-world networks. Both algorithms have a few control parameters and the parameter value is fixed after a number of trial-and-errors. Systematic parameter setting and theoretical analysis of learning process remain in future problems. Also practical applications should be considered in the future.

References

Study of LDPC Codes Built on Scale-Free Networks

Xia Zheng, Francis C. M. Lau and Chi K. Tse

Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
http://chaos.eie.polyu.edu.hk

Abstract—In this paper, we investigate, under an additive white Gaussian noise environment, the complexity of low-density parity-check (LDPC) codes built on scale-free (SF) networks and compare it with that of the best known LDPC codes. We compare the average number of connections in the nodes, which is a crucial factor in determining the hardware complexity of the LDPC encoder/decoder. Using the Gaussian approximation density evolution algorithm, we show that LDPC codes built on SF networks have a reduction of 4.7%–8.2% on the average number of node connections while achieving similar error-correcting capacity as the best known LDPC codes.

1. Introduction

Low-density parity-check (LDPC) codes were first proposed by Gallager in the early 1960s [1]. Due to the intense computational requirement, LDPC codes did not gain much intention in those days. With the substantial advancement of computational techniques in the last decade, the requirement of LDPC codes can be easily fulfilled today. Hence, LDPC codes have been “re-discovered” recently for in-depth investigations.

The original LDPC codes have a regular structure, meaning that all nodes of the same type have the same degree. By using bipartite graphs to represent the LDPC codes, improved codes (irregular LDPC codes) have been designed with irregular graphs [2]. It has also been shown theoretically that irregular LDPC codes can approach Shannon’s limit more closely compared to regular ones [3].

To optimize the LDPC codes, researchers usually refer to the threshold phenomenon. Given a code having an infinite block length and specified by the degree distributions of variable and check nodes. In theory, under a given channel condition, the best achievable performance of the code is determined by a certain “threshold” value. The larger the threshold value, the better the theoretical performance of the code. For example, under a binary erasure channel (BEC) and an additive white Gaussian noise (AWGN) channel, the threshold can be regarded as the maximum erasure probability and the maximum noise standard deviation, respectively, below which error-free communication can always be achieved.

To help find good codes with large threshold, a powerful tool, known as “density evolution (DE)”, has been proposed [3]–[4]. Although DE can compute the exact threshold of a code with arbitrarily small error probability, it is very complex. To reduce the computational complexity, a Gaussian approximation density evolution (GA-DE) algorithm has been proposed for an AWGN channel [4]. The GA-DE algorithm is much simpler compared to DE but the computed threshold is only an approximation to and is smaller than the actual one. Yet, even with the GA-DE algorithm, optimizing both the variable-node and the check-node degree distributions at a given code rate is a challenging task. Note also that in a practical environment, the code length is always finite.

The complexity of the encoder/decoder is another major concern to engineers who need to implement the system. While the variable-node and the check-node degree distributions are important factors that determine the best performance of a particular code, the average number of connections between the variable nodes and the check nodes is crucial in determining the interconnect complexity and the memory size and of the decoder as well as the amount of computations performed by the decoder [5]. Decreasing the average number of connections is a direct mechanism of reducing the complexity of the encoder/decoder. Alternatively, the decoding algorithm can be modified to reduce the decoder complexity at the expense of the capacity of codes [6, 7].

In recent years, complex networks have been studied across many fields of science including computer networks, biological networks and social networks. Complex networks are structures consisting of nodes interconnected by edges. In particular, for complex networks with a scale-free (SF) property, the distribution of node connections follows a power-law relationship, i.e., \( \Pr(k) \propto k^{-\gamma} \) where \( \Pr(k) \) denotes the probability that a node has \( k \) connections and \( \gamma \) is the characteristic exponent. Scale-free networks also have ultra-small diameters compared to other types of networks with the same number of nodes and connections [8]. Recently, it has been reported that irregular LDPC codes built on SF networks have very good error-correction capability under a BEC channel [9].

In this paper, we investigate, under an AWGN environment, the complexity of LDPC codes built on scale-free networks and compare it with that of the best known LDPC codes. We compare the average number of node degrees, which is a crucial factor in determining the hardware complexity of the LDPC encoder/decoder. In Section 2, we will briefly review some notations of the LDPC codes and describe the GA-DE algorithm. Section 3 describes the way to construct LDPC codes built on SF networks.
tion 4, we compare the complexity of our codes with other good codes known.

2. LDPC Codes and Gaussian Approximation Density Evolution

2.1. LDPC Codes

Low-density parity-check codes are linear block codes and can be expressed as solution vectors, denoted by \( x \), of the parity-check equation \( H x^T = 0^T \), where \( H \) is the parity-check matrix with sparse entries equaling 1. \( 0^T \) represents the all-zero vector, and the superscript \( T \) denotes the transpose of the matrix/vector. In the bipartite graph representation of LDPC codes, the code bits and the parity-check equations are denoted by two kinds of nodes, namely variable nodes and check nodes, respectively. Also, the variable nodes and check nodes are connected by edges, which are governed by the entries in the sparse parity-check matrix.

For each node, the number of edges connected is called the “degree” of the node. If all nodes of the same type have the same degree, the LDPC codes are regular; otherwise, the codes are irregular. For a given distribution pair \( (\lambda, \rho) \) of a LDPC ensemble,

\[
\lambda(x) := \sum_{i=2}^{d_v} \lambda_i x^{i-1} \quad \text{and} \quad \rho(x) := \sum_{i=2}^{d_c} \rho_i x^{i-1}
\]

specifies, respectively, the variable-node and check-node degree distributions. Also, \( d_v \) is the maximum variable-node degree and \( d_c \) denotes the maximum check-node degree. Moreover, the coefficients \( \lambda_i \) and \( \rho_i \), respectively, represent the fraction of edges connected to the variable and check nodes with degree \( k \). Based on the degree distributions, the code rate of the system, denoted by \( R \), can be obtained using

\[
R = 1 - \int_0^1 \rho(x) dx / \int_0^1 \lambda(x) dx.
\]

Figure 1 shows an example of \( (10, 5) \) irregular LDPC code. The \( (10, 5) \) code indicates that there are 10 variable nodes and 5 check nodes in the bipartite graph. It is readily shown that the degree distributions of the variable nodes and check nodes are, respectively, \( \lambda(x) = \frac{14}{23} x + \frac{9}{23} x^2 \) and \( \rho(x) = \frac{3}{23} x^2 + \frac{20}{23} x^4 \). Thus the code rate of the given \( (10, 5) \) irregular LDPC code is 0.5.

2.2. Gaussian Approximation Density Evolution

Each coded bit, denoted by \( b \), is represented by a signal with amplitude \((-1)^b\) for transmission through the channel. Without loss of generality, suppose an all-zero codeword has been sent. At the receiving side, assume that the transmitted signal is corrupted by independent and identically distributed (i.i.d.) Gaussian random variables with zero mean and variance (noise power) denoted by \( \sigma^2 \). Suppose a message-passing algorithm based on the log-likelihood ratio (LLR) is used in the iterative decoder [3]. Density evolution has been a powerful tool to track the distributions of the output messages at the variable nodes and the check nodes. However, the algorithm is very complicated. Chung et al. [4], approximating the probability density functions (pdfs) of the output messages at both variable nodes and check nodes by Gaussian functions and making use of the symmetry condition of all messages, i.e., \( f(x) = f(-x) \exp(x) \), where \( f(x) \) is the pdf of the LLR message, have simplified the complex DE into iterative equations that compute the means of the output message of the variable nodes and check nodes. In particular, the equation for the mean of the output message of a check node is given as

\[
m_{i-1} = \sum_{j=2}^{d_c} \rho_j \psi^{-1} \left( 1 - \left\lfloor \frac{\psi^{-1}(j-I^{-1})}{j-1} \right\rfloor \right)
\]

where \( m_{i-1} \) denotes the mean of the output message of a check node at the \( I \) iteration. Also, in (2),

\[
z_{i-1} = 1 - \sum_{i=2}^{d_c} \lambda_i \psi(m_0 + (i-1) m_{i-1})
\]

and

\[
\psi(x) = \begin{cases} 
\frac{1 - \frac{1}{\sqrt{4\pi}} \int_{-\infty}^{x} \tanh \frac{u}{2} \exp(-\frac{(u-x)^2}{4}) du}{1} & \text{if } x > 0 \\
1 & \text{otherwise}
\end{cases}
\]

with \( m_0 = 2/\sigma^2 \) and \( m^{(0)} = 0 \). Moreover, it has been shown that there exists a threshold \( \sigma^* \) below which \( m^{(I)} \) converges to infinity as \( I \) tends to infinity. This threshold, derived based on the GA-DE algorithm, is much easier to compute compared to the one based on the original DE. Like the threshold computed using DE, the larger the GA-DE threshold value, the better the achievable performance of the LDPC code with such degree distributions.

The discussion in this section indicates that the best achievable performance of a given LDPC code depends on
the distributions of the degrees of the variable nodes and check nodes. Optimization (maximization) of the threshold has been carried out and the degree distributions of some good codes have already been found [10].

3. Construction of LDPC Codes Based on Scale-Free Networks

Given a LDPC code that is corrupted by noise. The message-passing algorithm employed by the LDPC decoders allows the exchange of information between the variable nodes and check nodes during each iteration. As the iteration process goes on, the information generated by each variable node should eventually spread to all other check nodes and variable nodes. To visualize the flow of messages among the variable nodes, we remove the check nodes and construct a complex network using only the variable nodes. Moreover, the two variable nodes \( v_i \) and \( v_j \) are connected in the complex network only if they are connected to the same check node in the original bipartite graph. Figure 2 shows the complex network formed based on the LDPC code in Fig. 1.

Assuming that the total number of connections is the same, it is well-known that complex networks with SF property have smaller average path lengths compared to regular networks. While for the same average path length, complex networks with SF property have smaller number of connections compared to regular networks.

Consider the complex network formed by the variable nodes of a LDPC code. Based on the SF property of complex networks, if the variable nodes are interconnected with their degrees following a power-law distribution, the average path length between the nodes is smaller. This should enhance the exchange of information among the nodes and the original LDPC code has a higher potential of achieving good performance.

In practice, the variable nodes are connected to the check nodes. Fortunately, it has been proven that if the degree distribution of the variable nodes in the bipartite graph follows a power-law distribution, the degree distribution of the complex variable-node network formed also follows a power-law with the same exponent [11]. Moreover, if the exponent is strictly greater than 2, then the diameter of the network (i.e., the largest distance between any two nodes) has been shown equal to \( O(\log(N)) \) [11] or \( O(\log(\log(N))) \) [8] where \( N \) is the number of nodes.

We denote the fraction of variable nodes and check nodes with degree \( k \) by \( \Pr_v(k) \) and \( \Pr_c(k) \), respectively. To construct a LDPC code with the “scale-free variable-node-degree” property, we choose the fraction of variable nodes with degree \( k \) according to a power-law function, i.e., \( \Pr_v(k) \propto k^{-\gamma} \), where \( \gamma \) is the characteristic exponent for the variable-node degree. Then the variable-node degree distribution can be readily shown equal to

\[
\lambda(x) = \sum_{k=2}^{d_t} \frac{k^{1-\gamma}}{\sum_{i=2}^{\infty} i^{1-\gamma}} x^{k-1}.
\]

As for the check nodes, their degree distributions are also found to affect the performance of the LDPC codes to some extent. We have performed some simulations and find that LDPC codes with check-node degree following a Poisson distribution provides comparatively good results. This is also consistent with those reported in the literature [10]. In other words, the fraction of check nodes of degree \( k \) is selected according to \( \Pr_c(k) = \frac{\mu^e}{k!} \), where \( \mu \) is the average number of degree of a check node, and

\[
\rho(x) = \sum_{k=2}^{d_t} \frac{\mu^e}{k^{1-\gamma} \sum_{j=2}^{\infty} j^{1-\gamma}} x^{k-1}.
\]

Note that if the variable-node degrees, the maximum check-node degree and the code rate are fixed, the parameter \( \mu \) is easily determined from (1).

4. Results and Discussions

In this section, we compare the complexity of the LDPC codes built on SF networks with that of other best-known LDPC codes [10]. We assume a code rate of 0.5 and an AWGN channel. Table 1 presents the parameters used and the results. It can be observed that under all cases considered, the average number of connections for the codes from scale-free networks is 4.7%–8.2% smaller compared to that for other best-known LDPC codes. Note that the threshold values for the LDPC codes built on SF networks and for other LDPC codes, which are evaluated using GA-DE, are similar.

Finally, two codes are selected randomly from the LDPC code ensemble built on SF networks. The first one has a block length of 1000 and maximum variable-node degree 15 while the other one has a block length of 10000 and maximum variable-node degree 20. Figure 3 shows the degree distributions of the complex variable-node networks of the two randomly selected codes. It can be observed that the degree distributions are heavily tailed and look like
Table 1: Comparison of average number of connections ($<k>$) and threshold value ($\sigma^*$) evaluated using GA-DE between some best-known LDPC codes [10] and LDPC codes built on scale-free networks.

<table>
<thead>
<tr>
<th>Common Parameters</th>
<th>Optimized Codes in [10]</th>
<th>Codes Constructed on Scale-Free Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_v$</td>
<td>$d_c$</td>
<td>$&lt;k&gt;$</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\mu$</td>
<td>$&lt;k&gt;$</td>
</tr>
<tr>
<td>$%$ reduction in $&lt;k&gt;$</td>
<td>$\sigma^*$</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>4.0087</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>4.1638</td>
</tr>
<tr>
<td>30</td>
<td>23</td>
<td>4.4963</td>
</tr>
<tr>
<td>50</td>
<td>24</td>
<td>5.0765</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>0.9440</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>0.9460</td>
</tr>
<tr>
<td>30</td>
<td>23</td>
<td>0.9481</td>
</tr>
<tr>
<td>50</td>
<td>24</td>
<td>0.9523</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$%$ reduction in $&lt;k&gt;$</th>
<th>$\sigma^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2%</td>
<td>0.9427</td>
</tr>
<tr>
<td>5.5%</td>
<td>0.9487</td>
</tr>
<tr>
<td>4.7%</td>
<td>0.9542</td>
</tr>
<tr>
<td>7.3%</td>
<td>0.9580</td>
</tr>
</tbody>
</table>

Figure 3: The degree distributions of the complex variable-node networks of the two randomly selected codes. Left hand side: block length 1000 and maximum variable-node degree 15; right hand side: block length 10000 and maximum variable-node degree 20.

power-law except for the first few nodes with very low degrees. When the code length becomes longer, the scale-free property gets more prominent.

5. Conclusion

In this paper, we present a technique to construct LDPC codes with scale-free-network property. We also analyze their complexity and find that such codes have smaller average number of connections (hence lower hardware complexity) compared to other best-known LDPC codes with similar error-correcting capability. In future, we intend to further investigate the effect of other attributes of complex networks, e.g., clustering coefficient and mixing patterns, on the complexity of the LDPC codes built on SF networks.

Acknowledgments

This work was supported in part by a research grant (APG57) provided by The Hong Kong Polytechnic University.

References


Graph-based criteria for synchronization of diffusively coupled oscillators

Igor Belykh1, Martin Hasler2, and Vladimir Belykh3

1 Department of Mathematics and Statistics, Georgia State University, 30 Pryor Street, Atlanta, GA 30303, USA
2 School of Computer and Communication Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Station 14, 1015 Lausanne, Switzerland
3 Department of Mathematics, Volga State Academy, 5, Nesterov st., Nizhny Novgorod, 603 600 Russia

Abstract—We extend the connection graph stability method for complete synchronization, originally developed for symmetrically coupled networks, to the asymmetrical case. First, we study synchronization in asymmetrically connected networks with node balance, the property that the sum of the coupling coefficients of all edges directed to a node equals the sum of the coupling coefficients of all the edges directed outward from the node. We obtain the same criterion as for the network with a symmetrized connection matrix, provided that the condition of node balance is satisfied. Then, we derive a general synchronization criterion for arbitrary asymmetrical networks. The criterion is obtained by symmetrizing the connection graph and associating a weight to the path between any two nodes.

1. Introduction and problem statement

During the last few decades the study of the networks of coupled nonlinear dynamical systems has generated a rapidly growing interest in physics, biology and other fields of science and technology. The simplest and most striking cooperation between dynamical systems is their synchronization. The strongest form of synchrony in chaotic systems is complete synchronization when all oscillators of the network acquire identical chaotic behaviors. For a review on chaos synchronization, the reader is addressed to, e.g., [1, 2].

Complete synchronization in undirected and directed networks of linearly coupled limit-cycle and chaotic oscillators has received much attention (see [3, 4, 5, 6, 7, 8, 9, 10, 11] for a sampling of this large field). These studies show that both local and global stabilities of complete synchronization depend on the eigenvalues of the Laplacian connectivity matrix.

In a recent paper [12], we proved that the synchronization condition can also be derived from graph theoretical quantities using the connection graph method. The main step of the method is to establish a bound on the total length of all paths passing through an edge on the connection graph. This approach was originally developed for undirected graphs and applied to global synchronization in complex networks [13]. More recently, we showed how the method could be applied to directed networks [14].

The purpose of this paper is to present and review our recent results [14, 15] and proceed with the application of the connection graph stability method to the study of global synchronization in directed networks.

We consider a network of $n$ linearly coupled oscillators. We suppose that the individual dynamical systems are all identical, even though our results could be generalized to slightly differing systems. The equations of motion read

$$\dot{x}_i = F(x_i) + \sum_{k=1}^{n} c_{ik}(t)P_{x_k}, \quad i = 1, ..., n, \quad (1)$$

where $x_i \in \mathbb{R}^l$, the dynamical law of the individual system is expressed by $F : \mathbb{R}^l \rightarrow \mathbb{R}^l$ and $P$ is a projection operator that selects the components of $x_i$ that are involved in the interaction between the individual dynamical systems. They can range from a single component to all $l$ components. For clarity, we shall consider a vector version of the coupling when the first $s$ components are involved, and $P = \text{diag}(p_1, p_2, ..., p_l)$, where $p_h \equiv 1$, $h = 1, 2, ..., s$ and $p_h = 0$ for $h = s + 1, ..., l$.

The interaction is assumed to be of diffusive nature (on an arbitrary coupling graph). The coupling matrix $C$ is assumed to have nonnegative off-diagonal elements and zero row-sums:

$$c_{ik} \geq 0 \text{ for } i \neq k \text{ and } \sum_{k=1}^{n} c_{ik} = 0 \text{ for } i = 1, ..., n. \quad (2)$$

Matrix $C$ is assumed to be asymmetric, therefore it defines a directed (nonreciprocal) graph $G$ with $n$ vertices and $m$ edges. The vertices of the graph correspond to the individual oscillators, and the graph has an edge between node $i$ and node $j$ if at least one of the two coupling coefficients $c_{ij}$ and $c_{ji}$ is non-zero. To allow complete synchronization of all the oscillators, the graph is assumed to be connected.

The completely synchronous state of the system (1) is defined by the linear invariant manifold $D = \{x_1 = x_2 = ... = x_n\}$, often called the synchronization manifold. Our main objective is to obtain conditions of global asymptotic stability of synchronization in the system (1). We want to determine threshold values for the coupling strength required for synchronization, and to reveal their dependence on the network topology.
2. Asymmetrically coupled networks

In contrast to symmetrically coupled networks, where any connection graph configuration allows synchronization of all the nodes, synchrony in asymmetrically coupled networks is only possible if there is at least one node which directly or indirectly influences all the others. In terms of the connection graph, this amounts to the existence of a uniformly directed tree involving all the vertices. A star-coupled network where secondary nodes drive the hub is a counter example, where such a tree does not exist and synchronization is impossible.

It is worth noticing that the connections of node $i$ with the other nodes of the graph are defined by $i$th row and $i$th column elements of the matrix $C$. Connectivity matrices with both zero row and column sums correspond to node balanced networks, satisfying the property that all nodes in the network have equal input and output weight sums. We will first derive the synchronization criterion for such node balanced networks and then generalize it for arbitrary unbalanced coupling topologies.

Before we proceed with the study of directed networks, we should impose the following constraint on the dynamics of the coupled system (1).

2.1. Main hypothesis

**Assumption 1.** There exist a parameter $a > 0$ and a matrix $H = \text{diag}(h_1, ..., h_s, \bar{H})$, where $h_i = 1$ for $i = 1, ..., s$ and $\bar{H}$ is positive definite

such that the quadratic form defined by $H$ is a Lyapunov function for all the auxiliary linear systems (varying $x \in B_1$)

$$
\dot{\xi} = \frac{\partial F}{\partial x}(x)\xi - aP\xi
$$

simultaneously. Equivalently, all matrices

$$
H\left(\frac{\partial F}{\partial x}(x) - aP\right) + \left(\frac{\partial F}{\partial x}(x) - aP\right)^T H
$$

must be negative definite.

This constraint basically requires that the individual dynamical systems can be stabilized by adding a diagonal term for each state component that is involved in the interaction. In other words, we assume that there exists a critical value $a^*$, sufficient to make the equilibrium state $O$ of the auxiliary system (3) globally stable.

Assumption 1 is closely related to the requirement that the network (1) composed of two unidirectionally coupled systems globally synchronizes when the coupling $c_{ij}$ exceeds the critical value $a$. Many networks of linearly coupled limit-cycle and chaotic oscillator exhibit global synchronization such that Assumption 1 is satisfied. It was proved for coupled Lorenz systems [12], Chua circuits, Hindmarsh-Rose neuron models [13], etc. For example, for two unidirectionally $x$-coupled Lorenz oscillators:

$$
\begin{align*}
\dot{x}_1 &= \sigma(y_1 - x_1) + c_{12}(x_2 - x_1); \\
\dot{y}_1 &= r x_1 - y_1 - x_1 z_1; \\
\dot{z}_1 &= -b z_1 + x_1 y_1; \\
\dot{x}_2 &= \sigma(y_2 - x_2) \\
\dot{y}_2 &= r x_2 - y_2 - x_2 z_2 \\
\dot{z}_2 &= -b z_2 + x_2 y_2
\end{align*}
$$

(5)

Assumption 1 is true, and the bound for the synchronization coupling threshold is calculated as follows[12] $a = c_{12}^* = \frac{h_{12} + (r - 1)}{16(0 - 1)} - \sigma$.

2.2. Directed networks with node balance

It turns out that the graph-based stability criterion for asymmetrically but node balanced networks is identical to the stability condition for symmetrized networks where each directed edge is replaced by an undirected edge of half the coupling strength. Therefore, the connection graph stability method [12] can be directly applied to this class of directed graphs.

**Theorem 1** [14]. Consider the symmetrized connectivity matrix

$$
E = [\varepsilon_{ik}]: \quad \begin{cases} 
\varepsilon_{ik} = \varepsilon_{ki} = \frac{1}{2}(c_{ik} + c_{ki}), & \text{for } k \neq i \\
\varepsilon_{ii} = -\frac{1}{2} \sum_{k=1, k \neq i}^{m} (c_{ik} + c_{ki}), & \text{for } k = i
\end{cases}
$$

and let the original connectivity matrix $C$ be asymmetrical but have zero column sums (property of node balance).

Then, an upper bound for global synchronization in the directed network (1) with an asymmetrical connectivity matrix $C$ is calculated as follows:

$$
c_{ij} + c_{ji} \equiv \varepsilon_{ij} > \frac{a}{n} b_k(n, m) \quad \text{for } k = 1, ..., m,
$$

(7)

where at least one coefficient from the pair $(c_{ij}, c_{ji})$ is not zero and defines an edge on the directed connection graph; the mean coupling coefficient $\varepsilon_{ij} \equiv \varepsilon_{ij}$ defines edge $k$ on the undirected graph $E$ associated with the symmetrized matrix $E$; and $b_k(n, m) = \sum_{j=1, j \neq i}^{n} |(P_{ij})|$ is the sum of the lengths of all chosen paths $P_{ij}$ which pass through a given edge $k$ that belongs to the undirected graph. Here, $m$ is the number of edges of the undirected graph $E$ rather than the original directed graph.

Once again, the symmetrization operation amounts to replacing the edge directed from node $i$ to node $j$ by an undirected edge of half the coupling coefficient. In the case where there is an edge directed from node $i$ to node $j$ and another edge in the reverse direction, the pair of directed edges is replaced by an undirected edge with mean coupling coefficient. Finally, the calculation of $b_k$ along the symmetrized undirected graph $E$ gives us the stability condition for the asymmetrical network (1).

This calculation is straightforward within the framework of the connection graph method. To do so, we first choose a set of paths $P_{ij}$, $i = 1, ..., n$, $j > i$ (typically, the...
shortest paths), one for each pair of vertices \(i, j\), and determine their lengths \(P_{ij}\), the number of edges in each \(P_{ij}\). Then, for each edge \(k\) of the connection graph we calculate the sum \(b_k(n,m)\) of the lengths of all \(P_{ij}\) passing through \(k\). Our previous works [12, 13] give further details on a possible choice of paths and calculations of \(b_k(n,m)\) for different coupling configurations.

**Example: directed lattice on a torus.**

If all connection coefficients are equal, then node balance amounts to equal in- and out-degree of each node. Consider an example of such a network, the two-dimensional lattice on a torus with a uniformly directed node. Consider an example of such a network, the two-dimensional lattice on a torus with a uniformly directed node. Figure 1 (bottom). The torus is composed of \(n_{mer}\) and \(n_{par}\) oscillators in the meridian and parallel directions, respectively, \(n_{par} > n_{mer}\). Synchronization in this symmetrized network can be proved by calculating \(b_2\) (cf. Theorem 1). Due to the lattice structure of the network, we can simplify calculations by considering separately synchronization of oscillators in the meridian and parallel directions. It can be shown that this amounts to obtaining the synchronization conditions in rings of the sizes \(n_{par}\) and \(n_{mer}\). Considered together, these two conditions will then give the synchronization criterion for the entire network. More precisely, the maximum from \(b_2(n_{par})\) and \(b_2(n_{mer})\) will guarantee synchronization in the torus.

Synchronization in a ring of \(n\) coupled oscillators was already studied within the connection graph method [12]. It was shown that the graph quantity can be calculated as follows:

\[
C_k = \frac{N}{2} \sum_{i,j} C_{ij} \frac{1}{n^2}.
\]

Here, \(C_{ij}\) is the connection coefficient between node \(i\) and \(j\), \(N = n^2\) is the total number of nodes in the ring, and \(n\) is the number of nodes in the ring. The quantity \(C_k\) represents the total coupling strength in the network.

Theorem 1. The sufficient condition for the symmetrized torus network becomes

\[
c/2 > \begin{cases} 
\frac{a(n_{pr})^2}{2n^2} - \frac{1}{2n}, & \text{for odd } n_{pr} \\
\frac{a(n_{pr})^2}{2n^2} + \frac{1}{12}, & \text{for even } n_{pr}.
\end{cases}
\]

According to Theorem 1, this sufficient condition also guarantees synchronization in the original directed network Fig. 1 (top).

We have also calculated the eigenvalues numerically for specific torus network examples (for different \(n_{pr}\) and \(n_{mer}\)). For all these examples, the real parts of the eigenvalues of the directed and symmetrized undirected networks are the same. For example, for \(n_{pr} = 10\) and \(n_{mer} = 7\), the second largest eigenvalues of the directed and undirected networks are \(-0.1910 \pm 0.5878i\) and \(-0.1910\), respectively. It shows that our graph-based analysis correctly predicts the real relation between the synchronization properties of the two networks.

It must be emphasized that the real parts of the eigenvalues of directed but node balanced networks and their symmetrized analogs do not coincide in general. Typically, the symmetrized network is slightly more difficult to synchronize than the directed coupled one. This also supports our analytical results that the symmetrization applied to networks with node balance provides the sufficient condition that guarantees synchronization.

2.3. Connection graph method for arbitrary asymmetrical coupling

Here, we present a general framework for studying global complete synchronization in networks of dynamical systems with asymmetrical connections [15]. We derive our synchronization criterion in seven steps.

**Step 1.** Determine the "node unbalance" for each node \(C_j = \sum_{j \neq j} c_{ji}\). This quantity is defined to be the difference between the sum of connection coefficients of the outgoing edges and the sum of the connection coefficients of the incoming edges to the node.

**Step 2.** Symmetrize the connection graph by replacing the edge directed from node \(i\) to node \(j\) by an undirected edge of half the coupling coefficient \(c_{ij}/2\). In the case where there is an edge directed from node \(i\) to node \(j\) and another edge in the reverse direction, the pair of directed edges is replaced by an undirected edge with mean coupling coefficient \(c_{ij} = \frac{c_{ij} + c_{ji}}{2}\).

**Step 3.** To each edge of the symmetrized connection graph, we associate the quantity \(C_k\) defined by

\[
C_k = \begin{cases} 
\frac{C_i + C_j}{2n}, & \text{if } C_i + C_j < 0 \text{ and } k \text{ links } i \text{ and } j \\
0, & \text{otherwise}.
\end{cases}
\]
Step 4. Choose a path $P_{ij}$ between each pair of nodes. Usually, the shortest path is chosen.

Step 5. For each path $P_{ij}$ determine the mean node unbalance of the endnodes $i$ and $j$. To each path $P_{ij}$ associate its "length" $L(P_{ij})$ defined by

$$L(P_{ij}) = \begin{cases} |P_{ij}|, & \text{if } C_i^j + C_j^i < 0 \text{ and there is a link } k \text{ between } i \text{ and } j, \\ |P_{ij}|\chi(1 + \frac{C_i^j}{a}), & \text{otherwise} \end{cases}$$

(10)

where $C_i^j = \frac{C_j^i + C_i^j}{2}$ and $|P_{ij}|$ is the number of edges in $P_{ij}$. The function $\chi$ is the identity for positive and 0 for negative arguments.

Step 6. For each edge $k$ of the symmetrized connection graph determine the inequality

$$\epsilon_k + C_k > \frac{a}{b_k}, \text{ where } b_k = \sum_{j \in \epsilon(k \in P_{ij})} L(P_{ij}).$$

Step 7. Combine the inequalities either to describe the set of common values for all connection coefficients that guarantee global complete synchronization or to describe in general the set of connection coefficients vectors that guarantee synchronization if we allow for coefficients that vary from link to link. Finally, the bound for global synchronization in the symmetrized-and-weighted network holds also for the original asymmetrical network.

Remark 1. In the case where the directed connection graph is not a uniformly directed tree involving all nodes and complete synchronization of all the nodes is impossible, the condition for synchronization is simply impossible to satisfy.

Remark 2. If the "node unbalance" $C_j^i$ is zero for each node, we are back to node balanced networks. In this case, the seven-step process becomes Theorem 1.

We point the reader to [15] for further details of how to apply this general method to concrete asymmetrical networks.

3. Conclusions

We have developed a new graph-based method for synchronization in asymmetrical networks. The new ingredient of the method is the transformation of the directed connection graph into an undirected weighted graph. This is done by symmetrizing the graph and associating a weight to each edge of the undirected graph and to each path between any two nodes. This weight involves the "node unbalance" of the two nodes. This quantity is defined to be the difference between the sum of connection coefficients of the outgoing edges and the sum of the connection coefficients of the incoming edges to the node. As in the case of node-balanced networks, the synchronization criterion derived for this symmetrical network then guarantees synchronization in the asymmetrical directed network.

Acknowledgments

I.B. acknowledges the financial support of the Georgia State University Research Initiation Program (Grant FY07). M.H. acknowledges the support of the Swiss National Science Foundation through Grant N 200021-112081/1. V.B. acknowledges the support from the RFFI-NWO (Grant No 047.011.2004.004).

References


Tracking Cost of Mobile Agents in Ad Hoc Networks

Ryo Nakanishi, Shota Kawakami, Yoshikuni Onozato

†Department of Computer Science, Gunma University,
1-5-1 Tenjin-cho,Kiryu Gunma 376-8515, JAPAN
Email: kawakami@nttl.cs.gunma-u.ac.jp

Abstract—Use of mobile agents in ad hoc networks is now widely studied issue. For effective management of the mobile agent system, the location management is necessary. This paper concentrates on the functional extension of the path proxies scheme, which is one of the most basic and traditional location schemes for mobile agents, to be used in ad hoc network environment.

1. Introduction

An ad hoc network is a self-configured and infrastructure-less network connected by wireless signal transmission. Communications between long distance, over the 1-hop transmitting range, is supported by the multi-hop packet relay. All the terminals (often called "nodes") are supposed to be a mobile device and be equipped with limited energy supply (e.g. battery cell). A mobile agent, generally recognised, is a kind of software module that can autonomously migrate over the computer network and fulfill the tasks given by the user instead. Such mobile agent technology is now expected to be the foundation of the advance remote computing. The most important factor about mobile agents is during the agent work time, the user node does not need to be connected to the network, and can be asleep. Mobile agents are, in some forms, already proposed to use in ad hoc networks and relative works[1] are ongoing as well. This research also focuses attention on that mobile agent feature matches nodes’ demand in ad hoc environment.

2. Related Work

2.1. Location Management

Communication among agents subsumes the mechanism of locating. Moreover from a different viewpoint, the location mechanism is also prerequisite for further actions, such as to stop the agent activity, or to impose additional tasks. Hence location management (mechanism) is an essential issue in any mobile agent systems.

The core issue of the location mechanism is that an agent system is a highly dynamic open source system, where the number of agents varies considerably over time as new agents are born and existing ones terminated. Since mobile agents have autonomy and mobility, the whole system must be aware of each agent’s location at any time or be able to locate them within a tolerant time; however, too much to-locate-purpose information may cause critical overhead of traffic which invites defect to the system itself.

2.1.1. Preordained Migration Type

With an assumption that an agent migrates only along a pre-ordered path (e.g. by the user or system), it is easy to find the target agent by probing the different places its path. This probing process can be done either sequentially (i.e. checking single node each time), or parallel (i.e. probing more than one node each time). Throughout the probing process, some additional information that could be collected from the nodes in the path, for example whether the target has visited or not yet and/or the average migration interval, may be helpful for the further effectiveness of the search process. Example algorithms of this type include binary search (sequential) and ternary search (parallel).

2.1.2. Autonomous Migration Type

On the other hand, a fully autonomous agent does not have any paths known in advance. Instead of the unpractical aforesaid approach, following are known widely.

With an assumption that an agent migrates only along a pre-ordered path (e.g. by the user or system), it is easy to find the target agent by probing the different places its path. This probing process can be done either sequentially (i.e. checking single node each time), or parallel (i.e. probing more than one node each time). Throughout the probing process, some additional information that could be collected from the nodes in the path, for example whether the target has visited or not yet and/or the average migration interval, may be helpful for the further effectiveness of the search process. Example algorithms of this type include binary search (sequential) and ternary search (parallel).

If the inquiries are sent parallel (i.e. more than one inquiry at the same time), the result can be obtained simply faster. However, this scheme still depends on the same assumption of no migration during search period, unless the extreme case that inquiries are broadcasted to all of the places simultaneously, is adopted (of course this causes excessively harmful traffic to the network system). Furthermore, this approach is equal with the sequential approach in the message volume or slightly worse (more) than it.
3. Proposed Scheme

3.1. An Overview of the Proposed Scheme

The scheme this paper proposes is based on several assumptions. Such that each single node has their own neighbour table. This is provided with periodic update by using a beacon-like communication around them. Another is that each node has the same transmitting range (homogeneous) which is a widely adopted assumption in ad hoc network research. Also assuming that each node does not have any physical location device like GPS, thus they can not know their position whichever absolutely or relatively, but only the connectivity (graph) is obtainable information for each node.

To realise all aspects above, the solution this paper proposes is basically a greedy search with the tree-structure of the local network. Figure 1 shows brief idea of this mechanism.

First of all, each node of the tree represents the node in the network and links among them as well. In other words, the local network topology is copied into the tree-structure. The tree is only a known part of the network at the time of searching. The network topology is basically unknown for the tracking agent, but the topology is gradually being unveiled through the searching process. The tree is subgraph of the network graph then. The top of the tree is the last proxy node of path proxies, which has been reached from the tracking source. The tree is constructed without a flooding but done gradually and partially by referring a neighbour node table each node preserves. Nodes at the same level (same distance from the top node) of the tree form a set of nodes $D_i$ ($i = 1, 2, 3, \cdots$) where $i$ represents the distance (hops) from the top, and the searching (visiting every single node in a set) is executed set by set (namely $D_1, D_2, D_3, \cdots$). While the searching (visiting nodes), the tree is slightly extended as the newer part of the network is uncovered.

Furthermore, for the cost reduction purpose, every time the searching objects goes next level (between $D_i$ and $D_{i+1}$) the scheme sorts the all nodes at next level to the optimal order to visit (the sort algorithm will be discussed later).

A sequence of searching actions continues until any newer proxy is found.

3.2. The Sort Algorithm

This sort algorithm is invoked as a subroutine in the main algorithm, each time the searching proceeds to the next level of the tree. The purpose of this algorithm is to sort the nodes at next level in a particular order that the tracking agent takes as few hops as possible to visits the nodes. The algorithm aims avoidance of detouring and making effective use of potential links among the nodes.

Instead of discussing the concrete mechanism of the sort algorithm, the outline should be mentioned for the sake of convenience. Following steps are summarized and give only very brief idea of the mechanism.

step1 List up the object nodes (at the bottom level and not visited yet) from the tree, including duplication.

step2 Classify the nodes into subsets according to the parent node and pick up overlapped nodes between more than one subset as multi-parent nodes.

step3 For each multi-parent node, couple the both parents together (called parent-pair).

step4 Make one un-overlapped string from given parent-pairs (called parent-order).

step5 Sort the subsets according to the parent-order.

step6 Sort single-parent nodes in each subset according to the order of subsets and then insert multi-parent nodes between single-parent nodes of related two subsets accordingly.

step7 Select the order from ascending or descending, depending on the current query (TA) location.

4. Simulation

The main purpose of the simulation is to evaluate the performance of proposed scheme from the following two points of view: availability and cost effectiveness, with comparing those of the flooding scheme. The ratio of success to all tracking or flooding actions. The availability is defined as the ratio of success to all tracking or flooding actions. The tracking or searching is initiated from the node which has created a target agent. Whereas the number of sent messages is defined as the cost in flooding scheme, the cost of the proposed scheme is defined as the total hops finally required to reach the target agent.
4.1. Simulation Model

The simulation taken place in this research is plainly a random scenario Adjacency-Matrix-driven simulation. Generally recognised that an ad hoc network’s (and/or sometimes other forms of computer network’s) topology is often represented as a graph which has structure of vertices and edges in graph theory. An adjacency matrix is one of the expression forms of the graph and quite suitable to document the network topology and refer the graph itself. The adjacency matrix of one finite graph $G$ with $n$ vertices is represented as an $n \times n$ matrix where the non-diagonal entry $a_{ij}$ means there exists the number of edges between vertex $i$ and vertex $j$, and the diagonal entry $a_{ii}$ is the number of edges connected to and from vertex $i$ itself (called "loop"). In the simulation, the overall network topology is stored to one adjacency matrix, which ever the networks is connected or unconnected, and all simulation about searching and tracking are computed through the application of the matrix.

4.2. Results

4.2.1. Availability

The Figure 2 shows the mean availabilities of each scheme. Flooding scheme and proposed scheme show almost same tendency. The goal of this proposal is the cost reduction without losing availability.

The similar trend appears in different network mobility conditions, the only different parameter from previous ones is $P_m = 0.4$, which means that only 40% of the network nodes are mobile and the overall network mobility is restricted. The availability in these conditions tends to be similar and slightly higher in general.

4.2.2. Cost to WTT

The following Figures 3 to 6 show comparisons of costs by the flooding scheme and that by the proposed

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>network area</td>
<td>1000\times1000[m]</td>
</tr>
<tr>
<td>the number of nodes ($N$)</td>
<td>10,20,\ldots,300</td>
</tr>
<tr>
<td>transmitting range circle ($r$)</td>
<td>50,55,\ldots,200</td>
</tr>
<tr>
<td>mobility ratio ($P_m$)</td>
<td>0.4,0.6,0.8,1.0</td>
</tr>
<tr>
<td>speed of node movement</td>
<td>max/5[m/sec]</td>
</tr>
<tr>
<td>wait to track (max)</td>
<td>180[sec]</td>
</tr>
<tr>
<td>wait to track (interval)</td>
<td>10[sec]</td>
</tr>
<tr>
<td>the broadcast restricted in $\rho$ hops</td>
<td>10</td>
</tr>
<tr>
<td>required migration time of WA</td>
<td>10</td>
</tr>
</tbody>
</table>

Figure 2: Transition of each availability to WTT with fixed $N = 100, r = 120$ (sparse) and $P_m = 0.4, 0.6, 0.8, 1.0$.

Figure 3: Comparison of costs by flooding and proposed scheme when $N = 100$ and $r = 125$ with $P_m = 0.4, 0.6, 0.8, 1.0$.

Figure 4: Comparison of costs by flooding and proposed scheme when $N = 100$ and $r = 125$ with $P_m = 0.4, 0.6, 0.8, 1.0$. 

573
scheme in different network situations to WTT, where \( N = 100(r=120,125)200(r=90,95) \) respectively. All of these figures have the same horizontal axis WTT. As the overall tendency, in each series of figures, flooding is flat, which means the cost by flooding scheme is not affected by the time progress so much. On the other hand, proposed scheme is generally upward trend with an increase of WTT (more waiting time before tracking action). These overall trend is quite reasonable. On the other hand, the proposed scheme’s increasing trend sounds fair enough. This approach tries following along the path proxies and if the path proxy is broken, some additional cost will be expected; therefore, WTT plays the decisive factor for the total cost.

There is another parameter \( p_m \) in the simulation model. The proposed scheme has its additional cost according to the dissolution of the path proxies, hence easily inferable that, the network mobility concerns the whole cost. Figures 3 to 6 prove its effect. In general, the more nodes are mobile, the more the cost are expected.

5. Conclusion

This paper proposed mobile agents’ tracking mechanism in ad hoc environment, which is extended from the path proxy mechanism for stationary networks. Since the ad hoc network topology is supposed to be dynamic, the tracking system must take into account the topological change, which in turns is a series of path proxy collapses. The core of proposed scheme is without persisting in following the path proxy obstinately, but flexibly using the affordable information left by the Target Agent and last in the network; thus, gained the high availability alike the flooding scheme and succeeded in the cost reduction in general. However, in some sparse network environment, because of the difficulty of shortcut path creation, the cost can not be held in the better cost than just flooding approach but piles up.

References


Location-aided Energy-aware Routing Method for UWB-based Sensor Networks

Kyung Sup Kwak and Xizhi An

Graduate School of Information Technology and Telecommunications, Inha University
253 Yonghyun-dong, Nam-ku, Incheon 402-751, Korea
Email: kskwak@inha.ac.kr, anxizhi@hotmail.com

Abstract – In this paper, a new location-aided energy-aware routing method (eLAR) for sensor networks based on ultra-wideband (UWB) technique is proposed and evaluated. The eLAR makes use of the positioning capability of UWB and takes into account the energy efficiency in the network. By modeling the property of energy consumption, the eLAR routing algorithm is derived to search for energy-efficient routes that can support adequate QOS requirements. The complexity and scalability of this algorithm are also considered to make it more suitable for the operation of sensor networks. The simulation results have proved the advantages of this routing method.

1. Introduction

At present, sensor networks (SN) have already been penetrating people’s daily life, for example, in inventory, tracing/location, health-care, safety, productivity, and traffic-assist applications [1]. SNs consist of, possibly a large number of, tiny devices with sensing, computing, and communicating capabilities. Due to the limited energy and processing power of sensor, energy awareness is an essential issue in the SN design.

Impulse-radio ultra-wideband (IR UWB) technique [2] has many inherent properties suitable for SN applications. UWB systems have low complexity and low cost, use noise-like signal, are robust to multipath fading and jamming, and have very high time-domain resolution, allowing for location and tracking applications. For the energy consumption, UWB has the potential of efficient operation, since UWB has the lowest consumed energy per bit among different low-power RF technologies according to the simulation and measurement results [3].

There are many papers concerned with the routing scheme for SNs [4], however in this paper we want to emphasize the importance of utilizing location information in the route selection. Since SNs are usually set up in a fixed-wireless pattern or just have very low mobility, the location information of sensors can be obtained by some positioning methods. We adopt a SN scenario that applies UWB as physical layer in order to obtain its positioning capability [5]. We try to summarize a numerical expression of the relationship between energy consumption and route properties. Then a new routing metric concerned with energy can be devised to facilitate our routing method, which is a location-aided energy-aware routing scheme, named as “eLAR”.

This paper is organized as follows. Section 2 describes the system model used here. Section 3 analyzes the impact of multi-hop route on the energy consumption and discusses the realization of eLAR. Section 4 gives simulation results of eLAR performance. Finally, our conclusions are presented in Section 5.

2. System Model

The physical model is based on the Time-Hopping (TH) Pulse Position Modulation UWB [2]. The chip time is $T = 0.2$ ns and pulse repetition period $PRP = 280$, so the maximum data rate is $1/(T \times PRP) = 18$ Mbps. The transmit power of sensor node is limited so that the radio coverage radius reaches to about 24 meters (denoted as $C = 24$ m). All sensors have the same maximum transmit power and they transmit at the maximum power if they have data to send, since the power control is inefficient in UWB networks [6]. In addition, the statistical UWB indoor path-loss model [7] is used in our work.

The MAC layer utilizes the Dynamic Channel Coding (DCC) MAC protocol [8]. DCC-MAC fully exploits the advantage of IR UWB and is composed of two parts: one is DCC codec and the other is private MAC. The former performs the rate-compatible punctured convolutional codes and adapts coding rate according to the interference and channel conditions. The latter can resolve contentions by using TH sequences generated pseudo-randomly according to MAC address of node. It should be noticed that DCC equivalently realizes a kind of “soft” power control that distributes appropriate energy to data bits through variable coding gain.

In general, sensors measure specific conditions and transfer information to a data sink. This kind of data transfer is either periodical or event-driven and may have some quality of service (QOS) requirements, such as transfer delay and packet delivery ratio. For simplicity, UDP is applied to emulate the transport layer and the packet size is fixed as 128 bytes in this paper. Our purpose is to design an effective routing scheme that can provide adequate services to upper layers and at the same time consume as less as possible energy of hardware platform.

From the perspective of energy saving, sensor nodes usually have four states, that is, “Transmit”, “Receive”, “Idle” and “Sleep”, and each of them has a certain power level. For UWB devices, the target power consumption of circuit hardware may be $100$ mW [9]. In this paper, we focus on the power consumption of packet transmitting-
receiving procedure, so we simply set the “Transmit” power as 60 mW, the “Receive” power as 30 mW, the “Idle” power as 0 mW and the “Sleep” power as 0 mW.

3. Routing Scheme Design

3.1. Impact of Multi-hop Route

The route which packets pass through surely has great influence on the energy consumption and QOS issues. We define following items to evaluate these issues.

1) Total Energy Consumption per Packet Delivered \( (E_p) \). \( E_p \) is the ratio of the total energy consumed by nodes participating packet delivery over the number of packets successfully delivered during the same period.

2) End-to-end Packet Transfer Delay \( (T_d) \). \( T_d \) is the average end-to-end transfer delay of packets that are successfully delivered from sources to destinations.

3) Packet Delivery Ratio (PDR). PDR is the ratio of the number of packets successfully delivered over the total number of packets sent by sources.

To begin with, a line-style network topology as shown in Fig. 1 is studied. The Sensor transfers data packets to the data sink through either a direct (single-hop) connection or a multi-hop route. The number of hops is denoted as \( H \). The length of hop, that is, the distance between any two adjacent nodes, is the same as \( L \), and the distance between the Sensor and the Sink is \( D = H \times L \). Let \( L_R \) denote the length of route. In this case, \( L_R \) is also equal to \( D \).

Fig. 1. Line network topology. The distance between the Sensor and the Sink is \( D \). The length of each hop is the same as \( L \).

The impact of route properties, such as \( H \), \( L \) and \( D \) \( (L_R) \), can be learned directly by implementing simulation in the ns-2 [10]. This empirical method is reasonable, because the network performance depends on the average signal level in the long run and we can focus our attention on the effect of route properties. Simulation results are presented in Fig. 2.

As shown in Fig. 2 (a), curves of \( E_p \) seem to have a linear lower bound which is proportional to \( D \), if \( L \) is less than about 12 meters. If \( L \) is larger than 12 meters, \( E_p \) rises up abruptly with the increase of \( L \). When \( L \) is close to \( C \), \( E_p \) tends to be infinite, that is, reliable transmission is impossible. Moreover, if a route contains several long hops \( (L > 12 \text{ m}) \), the corresponding \( E_p \) is much higher than that of the route consisting of short hops for the same \( D \). This situation suggests that there would be a threshold of \( L \), which decides the energy efficiency of relay process.

The main factor behind this \( E_p \) difference is the mean path loss that is proportional to the propagation distance raised to some exponent. The transmitted signal could be attenuated greatly over long distance and thus the SNR level at the receiver would be too low to guarantee good reception. In order to correct bit errors or frame errors, lower-rate channel coding or retransmission can be applied, but inevitably the efficiency of data transmission is reduced and hence energy consumption is increased correspondingly. Meanwhile, transfer delay and packet loss as shown in Fig. 2 (b) and (c) are increased too.

It should also be noticed that, within a certain range of \( D \), the route with less \( H \) tends to have lower \( E_p \), if its \( L \) does not exceed 12 meters. For example, in Fig. 2 (a), the direct connection has the minimal \( E_p \) among all routes if \( D \) is within \( (0, 12 \text{ m}) \), and the 2-hop route shows its advantage if \( D \) falls in \( (12 \text{ m}, 24 \text{ m}) \). The reason is that the probability of transmission collisions rises up with the increase of \( H \) in a certain small range and more intermediate nodes mean more interference, processing and propagation, resulting in more energy consumption.

Note that the \( E_p \) curve and the \( T_d \) curve have the similar tendency of change due to route properties. Since the packet delay just is the time duration of packet being processed within the network, if the processing power is fixed, the energy consumed by packet delivery would be proportional to the delay time. The \( PDR \) curves shown in Fig. 2 (c) are also influenced by route properties in a similar way. If \( L \) is too large, \( PDR \) could not satisfy the requirement of upper layers. On the contrary, the route with short hops can ensure reliable data transfer. These findings implicate that an energy-efficient route could support adequate QOS requirements at the same time.

3.2. Energy Metric of Route

To quantify the impact of multi-hop route, we develop a location-aided approach to estimate the energy metric of route. It is assumed that every node knows positions (relative distances) of all nodes that compose the network.

Let \( L_R \) be the length of the route connecting one sensor and the sink. Then \( L_R \) is given by

\[
L_R = \sum_{i=1}^{R} L_i ,
\]

where \( \{ L_i \} \) is the length of hop \( i \) belonging to the route.

Then an optimal forward distance \( F \) is defined. If \( L_i \leq F \) for all \( i \) and \( D > F \), the lower bound of \( E_p \) is expressed as

\[
E_{min} = H \cdot E_{relay} + \lambda \cdot L_R,
\]

where \( \lambda \) is the proportion constant and \( E_{relay} \) is the energy consumed by relay node. Thus, the item \( \lambda \times L_R \) accounts for the increase of energy consumption due to the path loss of the route, and \( H \times E_{relay} \) is corresponding to the inherent processing load of relay nodes.

If there is a hop longer than \( F \), an additional increase in \( E_p \) should be taken into account. In a simple form, this addition resulted from each hop \( i \) is

\[
a_i = \begin{cases} 
1, & \text{ when } L_i \leq F \\
(C - F)(C - L_i), & \text{ when } F < L_i < C \\
0, & \text{ when } L_i \geq C
\end{cases}
\]

Finally, the energy metric of route is given by
$$E_p = E_{\text{ms}} \prod_{i=1}^{H} \alpha_i .$$

Note that if \( D \leq F \), the direct connection is preferred.

### 3.3. Routing Algorithm

Our new routing algorithm, called eLAR, originates from the Dijkstra’s Algorithm [11] and there is a bit modification. The network topology can be regarded as an undirected graph and each node is a vertex in the graph. The length (or cost) of a hop (or edge) is not directly used, but the overall \( E_p \) on the route containing that hop is evaluated. Every node in the network will by itself find the “shortest” paths to the sink, or other nodes if needed. Here, “shortest” means that the corresponding \( E_p \) is minimal. The following is a high-level algorithm for this approach.

1. \( V = \{ v_1, v_2, \ldots, v_N \} \); // set of nodes, \( N \) nodes in total
2. \( Y = \{ v_i \} \); // \( v_i \) is the source node
3. \( F = \Phi \); // min. loss tree (initially empty)
4. while (\( V \neq Y \) ) {
5. \( \text{select a node} \ v \ \text{from} \ V-Y, \ \text{that has the min energy loss} \)
6. \( \text{from} \ v_i, \text{using only nodes in} \ Y \text{as relay nodes;} \)
7. \( \text{add the new node} \ v \ \text{to} \ Y; \)
8. \( \text{add the hop that touches} \ v \ \text{to} \ F; \)
9. }

A routing table can be derived from the minimum loss tree at each node, and packets can be forwarded sequentially by nodes on the route without an extra route header. The eLAR scheme can be easily implemented as following steps:

1. The source (src) measures whether the destination node (dst) is in its near vicinity. If the distance between src and dst, \( D \), is not larger than \( F \), src directly transmits packets to dst, otherwise, src searches its routing table to find the next-hop node (nxt) that is on the minimum loss route to dst and then forwards packets to nxt.
2. The relay node (rly) checks the destination (dst) of each received packet. According to the distance between rly and dst, rly performs the same action as Step 1).

### 3.4. Algorithm Scalability and Simplification

Note that there are \( N-1 \) iterations and the number of operations per iteration is proportional to \( N \), so the computation required by the routing algorithm is \( O(N^2) \) in the worst case. Since SN nodes only have very limited capability of computing with very small memory, the algorithm may be hardly implemented in sensors. If the network scale is small, the eLAR algorithm could be feasible; however it would be too complex to implement in a large network scenario. In order to solve this problem, a simplified eLAR algorithm (eLARs) is proposed as follows.

The src performs eLAR algorithm only within the radius \( C \) to find the “shortest” paths to all its neighbors and estimate the corresponding \( E_p \). If the src-dst pair is not in communication range, the src measures the Most Forward Efficiency (MFE) among its neighbors in terms of forward progress in the direction to the final destination. The MFE is given as:

$$\text{MFE}^e = \max_{k: \text{src} \rightarrow \text{neighbor}} \left\{ \frac{d_k \cdot \cos \theta_k}{E_{p_k}} \right\},$$

where \( d_k \) is the distance between src and neighbor node \( k \), \( \theta_k \) is the angle of node \( k \) in terms of the src-dst direction. Hence, \( d_k \cdot \cos \theta_k \) denotes the forward progress of node \( k \) towards the dst. Therefore, the MFE accounts for selecting an exit node that has the highest forward efficiency, that is, packets can be forwarded as close to the dst as possible with a certain amount of energy consumption. Finally, the src can find the nxt corresponding to the exit node and delivery packets to it.

### 4. Simulation Results

The network simulator ns-2 [10] is utilized to evaluate the performance of eLAR/eLARs scheme and compare it with other routing protocol, e.g., AODV [12]. The eLAR parameters are as follows: \( C = 24 \) m, \( F = 11 \) m, \( \lambda = 0.005 \) and \( E_{\text{relay}} = 0.06 \). Other parameters take default values.
Fig. 3 shows the network scenario in which 48 nodes are randomly placed in an area of 50 m × 50 m and do not move. Sensor nodes are numbered from 1 to 47 and the data sink is especially marked as 0. Sensors are randomly picked as source nodes to transfer packets to the sink.

In the simulation, sensors #1, #3, #8, #12, #19, #29, #33 and #36 are selected and investigated. Simulation results are shown in Fig. 4. Fig. 4 (a) gives the comparison of $E_p$ between AODV and eLAR/eLARs. Only for the close connections #3, #8 and #29 can the performance of AODV be comparable with eLAR scheme. For the sensors #1, #12, #19, #33 and #36, the routes selected by eLAR/eLARs are obviously energy-efficient and have more advantages. Fig. 4 (b) and (c) demonstrate that the routes selected by eLAR/eLARs can provide fairly good QOS at the same time. Note that the energy consumption, transfer delay and delivery ratio of eLARs routes are very close to those of eLAR routes, so the simplification method in eLARs is proved to be effective. In some cases, the performance of eLARs is a little lower than eLAR, since after all eLAR can give an optimal solution throughout the network, but eLARs is much simpler and more suitable for sensors operating in a big SN.

We can see that AODV may not be suitable for SN applications, especially for long-distance connections. eLAR utilizes location information and simplifies the signaling procedure. eLAR can effectively deal with the contradiction between relay selection and transmission efficiency, since the energy metric of route ensures the result is energy-efficient and the routing algorithm provides the connectivity throughout the network.

5. Conclusions

This paper proposes a practical routing scheme, eLAR, in which the energy consumption of multi-hop route is taken into account and the relationship between energy and QOS issues is preliminarily discussed. eLAR algorithm is simplified effectively to be more suitable for sensors’ operation and can improve energy efficiency in different cases, because eLAR simplifies the distributed signaling procedure and selects direct connection or multi-hop route in a proper manner.

References

Context Acquisition via Multifarious Sensors in a Real World Home Ubiquitous Environment

Tatsuya Yamazaki

Universal City Group, Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
Email: yamazaki@nict.go.jp

Abstract—The National Institute of Information and Communications Technology of Japan has constructed a real world home environmental test bed, called the Ubiquitous Home, for home context-aware service experiments. Various kinds of sensors are installed in this test bed to collect human behavior and user context in real-life. Usually who, where and when are considered as main contextual information. To obtain the information of who, an RF-ID tag system are used. A floor pressure sensor system as well as the RF-ID tag system are used to obtain the information of where. A common NTP system is used for acquisition of the information of when. This paper shows the acquired data in experimental and real-life environments and how these data are used to construct context-aware services in the home environmental test bed. In addition, radio wave environmental data of the test bed are presented because they influence the performance of the RF-ID tag system.

1. Introduction

Ubiquitous computing [1] have been realized owing to the progress of wired and wireless home networking, sensor networks, networked appliances, mechanical and control engineering and computers themselves. Using these developed technologies, smart home projects have started at every corner of the world.

In 1995, the Welfare Techno Houses in Japan was constructed [3]. The concept of these experimental houses was to promote independence for elderly and disabled persons and to improve their quality of life.

The Aware Home [4] project is noteworthy among the smart home researches. In the Aware Home project, they built a three-story, 5040-square-foot home that functions as a living laboratory for interdisciplinary design, development and evaluation. The University of Texas at Arlington has conducted MavHome project [5]. The MavHome is a home environment that perceives the state of the home through sensors and intelligently acts upon the environment through controllers. Also, at the Massachusetts Institute of Technology, the House_n group [6] is working towards a vision where computer technology is ever-present, but in a more subtle way than often advocated in popular culture and even in engineering paper motivation sections. They want sensor-driven pervasive technologies to empower people with information that helps them make decisions, do not want to strip people of their sense of control over their environment.

Extended from Robotic Room 2 [7], Sensing Room was constructed at the University of Tokyo [8]. Although they say Sensing Room can measure detail of human daily behaviors in a long term, it seems that there is some limitation to imitate a real-life within a small room only. The sensing room of EasyLiving project at Microsoft Research [9] is similar to Sensing Room. In EasyLiving project, they aimed at developing prototype architecture and technologies for building intelligent environments.

The University of Florida's Mobile and Pervasive Computing Laboratory is developing programmable pervasive spaces in which a smart space exists as both a runtime environment and a software library [10]. The University of Sherbrooke of Canada constructed DOMUS laboratory, which is a new research pole on cognitive assistance in smarts homes and mobile computing [11]. As DOMUS laboratory is situated in the University of Sherbrooke, atmosphere is different from real-life in a sense.

Active Home is also located in a university, Information and Communications University of Korea and several experiments of context-aware services according to human behaviors are carried out [12]. UbiHome is another smart space in Korea, where automated control of lights and monitors according to user's situations are demonstrated, for example [13].

In several smart home projects, one purpose is to provide a service to (a) user(s) accordingly. To provide an appropriate service, contextual information, that is user context, should be considered. Although the definition of user context may change according to the situation or the service used, who, where and when are usually main contextual information. Although multifarious sensor collaboration is one solution to get the user context, few reports are presented on how to combine multifarious sensors.

The National Institute of Information and Communications Technology of Japan has constructed a real world home environmental test bed, called the Ubiquitous Home, for home context-aware service experiments. Various kinds of sensors are installed in this
2. The Ubiquitous Home

2.1. Paper Format

The Ubiquitous Home is an ICT housing test facility for the creation of useful new services for the home that will become possible by linking devices, sensors and appliances by means of a data networks. The Ubiquitous Home has a living room, dining-kitchen, study, bedroom, washroom and bathroom, these rooms comprising an apartment. In addition to the apartment, a Japanese-style room is provided as a living space for remotely living family members, such as a grandmother and/or a grandfather. Between the apartment and the Japanese-style room is a computer room called the Network Operating Center (NOC).

Above the ceiling of the Ubiquitous Home is a space where experimenters can work. In the space, three corridors called “cat walks” are established. These are used for walking, machine installation and cabling, respectively. Moreover, the floor of the Ubiquitous Home is structured for free access to 40 cm height.

The Ubiquitous Home is equipped with various types of sensors to monitor living human activities. Each room has cameras and microphones (Fig. 1) in the ceiling to gather video and audio information. To respect residential privacy in such ubiquitous camera and microphone environments is one of the important issues to be considered.

Floor pressure sensors (Fig. 2) installed throughout the flooring contain 18 cm by 18 cm binary detection units and are used to track residents or detect furniture positions.

Infra-red sensors (Fig. 3) installed at the top of the entrance door to each room and at foot positions in the kitchen and in the corridor detect infra-red sources literally, so can be used to detect human movement.

Two RFID systems are installed in the Ubiquitous Home. One is an active-type RFID system, the other is a passive-type RFID system. The former uses 315MHz waves, the latter the 2.45 GHz band. Active-type system scanners located over the ceiling of each room detect the entrance of RFID tags into the room. Passive-type system antennas are embedded inside the wall around each room entrance. When an RFID tag passes through the entrance, the antenna reads the information on the tag.

Four accelerometers or vibration sensors are attached to back of the bedroom floor in four corners. Although the role of the accelerometers is similar to the floor pressure sensors, they are more promising to detect human behaviors owing to their higher precision than the floor pressure sensors.

Displays (plasma display panels and liquid crystal displays) (Fig. 4) and speakers are installed throughout the Ubiquitous Home to provide residents with video and audio contents.
As the user context, location, time and personal identification are fundamental information. In the Ubiquitous Home, the place information can be detected by the active-type RFID tag and the floor pressure sensors. The active-type RFID tag can capture the user location as the level of room globally and the floor pressure sensors supplement more precise information. The time information can be provided a computer clock which is adjusted by a Network Time Protocol (NTP) server. This is requisite for combine multifarious sensor information. The personal identification can be obtained from the active-type RFID tag worn by the resident.

3. Wireless Environments in the Ubiquitous Home

As described in the above section, the RFID systems play an important role to collect the user context information. The RFID tag detection is not, however, so stable in the Ubiquitous Home. Therefore we measured the wireless environments in the Ubiquitous Home. Figure 5 shows the results, where the horizontal axes present the frequency and the vertical axes present the field strength in dBuV/m.

(a) and (b) show the relative relationship between the 315 MHz active-type RFID tag system and all other equipments and appliances. It is found that the RFID tag can be affected by other equipments that emit almost same frequency wave.

(c) and (d) show the relative relationship between the 2.4 GHz passive-type RFID tag system and all other equipments and appliances. In these range of frequencies, there are IEEE 802.11b and Bluetooth that can be used without any license. Therefore some detection was observed even if all equipments and appliances are turned off in the Ubiquitous Home. Since the Ubiquitous Home is built inside our research center, the waves from wireless LAN system used in other laboratories seem to be detected. It is found that there is possible confliction of used wave.

4. Conclusions

In this paper, utilization of multifarious sensors to detect the user context information is presented. The scheme was implanted in the Ubiquitous Home, a real world home environmental test bed. Especially, RFID tag systems and floor pressure sensors are utilized and they are synchronized by using an NTP server. Since the RFID tag systems are easily affected by other equipments and appliances, the wireless environment characteristics have been measured in the Ubiquitous Home. As the result, it is found there are influence or conflict from other equipments. Some countermeasures have to be considered against it.
Acknowledgments

The author would like to thank all contributors in the UKARI project for their support in constructing the Ubiquitous Home. The members of the UKARI project as of November in 2004 are Toshiba Corporation, NEC Corporation, Sanyo Electric Corporation, Oki Electric Industry Co., Ltd., Dai Nippon Printing Co., Ltd., Samsung Electronics Co., Ltd., Kyoto University, Doshisha University, Ritsumeikan University, Osaka Institute of Technology and National, University of Tokyo, Bukkyo University, Kochi Women's University, Wakayama University, Okayama University and National Institute of Information and Communications Technology.

References


Chaos-Based Generation of Optimal Spreading Sequences for DS-UWB Sensor Networks

G. Cimatti, R. Rovatti†, G. Setti‡

†DEIS - University of Bologna, Viale Risorgimento 2, 40136 Bologna - Italy
‡ENDIF - University of Ferrara, Via Saragat 1, 44100 Ferrara - Italy
All the authors are also with ARCES-University of Bologna, Via Toffano 2/2, 40125 Bologna - Italy
Email: {gcimatti, rrovatti, gsetti}@arces.unibo.it

Abstract—We verify the effectiveness of chaos-based sequence generation in reducing Multiple Access Interference (MAI) in Direct Sequence UWB Wireless-Sensor-Networks (WSNs). Extensive simulations have been performed and show that it is possible to increase the expected Bit-Rate (BR) at which each user may transmit given a certain Signal-to-Interference Ratio (SIR).

1. Introduction

Wireless Sensor Networks (WSN) are collections of active nodes whose position, distance, number and operating environment cannot be anticipated. As such, their functionality depends on the existence of a communication network capable of transferring information from one node to another and possibly to a central collecting unit.

Due to this structure, the physical wireless interface of a such a network must cope with the presence of a potentially large number of competitors for the communication channel as well as with the absence of a central authority managing the diversity needed to distinguish competing logical links.

A way to address this problem is to employ Direct Sequence Code Division Multiple Access (DS-CDMA) that distinguishes nodes by means of spreading codes that act as node signatures. Since no global structure or authority is assumed, spreading codes cannot be made orthogonal unless a very complex negotiation is initiated that risks to involve the whole network and thus must be avoided [1]. This causes Multiple Access Interference (MAI) that actually becomes the main reason of quality degradation when the number of nodes increases. To avoid frequency allocation problems (and thus cope with uncontrolled and unsupervised environments) DS-CDMA modulation may employ Ultra Wide Band (UWB) pulses designed to obey recent regulations on radio links with extremely low power spectral density that are allowed to superimpose already allocated bands.

Our idea is to use chaos-based generators to control the statistical features of the sequences that, in turn, interact with the pulse shape in determining the final performance figure. Monte-Carlo simulations have been performed to tune the chaos-based generator to the pulse shape and obtain an increase in BR when compared with traditional PN generation.

2. System Model

The signal generated by the $u$-th transmitter can be written as

$$s^u(t) = \sqrt{P_u} \sum_{h=\infty}^{\infty} b_h^u \sum_{k=0}^{N-1} y_k^u g(t - h T_b - k T_c),$$

where $b_h^u \in \{-1, 1\}$ are the transmitted bits, $y_k^u \in \{-1, 1\}$ represent the spreading symbols, $T_b$ is the bit period, $T_c$ is the chip period, the integer $N = T_b/T_c$ is the spreading factor, and $g(t)$ is the chip pulse. The factor $\sqrt{P_u}$ is the transmitted power and depends on $N$, on the pulse profile, on the sequence $\{y_k^u\}$ and it sets the energy $E_b^u$ that the transmitter puts in each bit. Conventional DS systems employ m- or Gold-sequences. Here we propose a chaos-based sequence generation method which allows the reduction of MAI. Spreading sequences with any exponential decay in correlation functions can be obtained from the quantization of a discrete time chaotic dynamical system dependent on the function $M(x)$ whose behavior can be attacked by means of the tools of statistical dynamics [2]. The family of maps we will use is defined as

$$M(x) = \begin{cases} \frac{x}{2^r} & \text{if } 0 \leq x < p \\ \frac{x + 1 - 2p}{2^r} & \text{if } p \leq x < 1/2 \\ \frac{x - 1 + 2p}{2^r} & \text{if } 1/2 \leq x < 1 - p \\ \frac{x + 2^r - 1}{2p} & \text{if } 1 - p \leq x \leq 1 \end{cases}$$

where $r \in [-1, 1]$ and $p = (1 + r)/4$ is the parameter used to set the breakpoints between adjacent segments. The quantization function $Q(x)$ is such that $y_j = -1$ if $x \leq 1/2$ and $y_j = +1$ otherwise. We are therefore guaranteed for $i \leq j \leq k \leq l$ we have $E_y[y_i^u y_j^v y_k^w y_l^v] = r^j-i$ and $E_y[y_i^u y_j^v y_k^w y_l^v] = r^{j-i+k-l}$, where the $E_y[\cdot]$ operator gives the expectation of its argument with respect to all the possible spreading sequences. Note that when $r = 0$ the corresponding map is able to generate independent and identically distributed (i.i.d.) spreading symbols. In [3] it has been proved, for a synchronous DS-CDMA system, that m- and Gold-sequences provide almost the same performance of i.i.d. sequences, so in the following we will refer only to i.i.d. The signal generated at the
transmitter is sent through the channel by means of an antenna that can be effectively modelled by an attenuation $A_{tx}$ (assumed equal for all the nodes) and a time-derivative [4]. Along the channel the signal suffers an attenuation $A_{ch}$ and a delay $\tau_u$ that depends on the traveled path that we here assume unique. It then arrives at the receiver antenna which imposes a further attenuation $A_{rx}$ and time-derivative. The signal arriving at the receiver due to the $u$-th transmitter can be written in terms of the second derivative of the chip original pulse \[5\] $g_{rx}(t) = (d^2/dt^2)g(t)$. Since we assume that receivers can eventually compensate delays that do not change in time, we are legitimate to set $\tau_u = 0$ for every $u$. Furthermore all the nodes have identical air interfaces and thus deliver the same energy per bit $E_b = E_n/\nu u$. We may concentrate on the reception of the 0-th bit, choose $A_{rx}A_{ch}A_{tx} = 1$ as a reference ideal conditions and get

\[
E_b = \int_{t_0}^{t_f} \left| \sqrt{P_u} \sum_{k=0}^{N-1} y^o_{k}(x-kT_c) \right|^2 dt,
\]

\[
P_u = \frac{\sum_{j=0}^{N-1} \sum_{k=0}^{N-1} y^o_{k} y^c_{j} G_{rx}((j-k)T_c)}{\sum_{j=0}^{N-1} \sum_{k=0}^{N-1} y^o_{k} y^c_{j} G_{rx}((j-k)T_c)},
\]

to the receiver, where we have assumed $g(-t) = g(t)$ and thus $g_{rx}(-t) = g_{rx}(t)$ to define $G_{rx}(x) = \int_{-\infty}^{\infty} g_{rx}(x-t) dt$ as the convolution of the chip pulse with itself. From the above equality the received power by the $u$-th node is

\[
E_b = \int_{t_0}^{t_f} \left| \sqrt{P_u} \sum_{k=0}^{N-1} y^o_{k}(t-kT_c) \right|^2 dt
\]

Note that when received pulses do not produce inter-channel interference (ICI) we have $G_{rx}(mT_c) = 0$ for every integer $m \neq 0$ and $G_{rx}(0) = \int_{-\infty}^{\infty} g_{rx}(x) dx$ that is the pulse energy. UWB pulses are, in general, non-ICI-free and $P_u$ takes into account the whole bit signal. In the following we will assume that $P_u$ is such that $E_b = 1$.

The receiver is fed with the sum of all the incoming signals plus the thermal noise $n(t)$, i.e. with

\[
E_b = \sum_{u=0}^{U-1} \sum_{h=-\infty}^{\infty} b^o_{h} y^o_{h}(t-hT_h)
\]

where $y^o_{h}(t) = \sqrt{P_u} \sum_{k=0}^{N-1} y^o_{k} g_{rx}(t-kT_c)$ is the waveform corresponding to the unmodulated spreading sequences at the receiver. We assume to use the simplest possible decoding, i.e. a correlate and dump receiver matched with the useful transmitter. Within this scheme we focus, without loss of generality, on the decoding of the 0-th bit of the 0-th transmitter and first compute

\[
X = C \int_{\infty}^{\infty} z(t) y^o_{rx}(t) dt
\]

for some positive constant $C$ and estimate $b^o_{h} = 1$ if $X > 0$ or $b^o_{h} = -1$ if $X < 0$.

Note that $X_u = x_u$ thanks to the factors $P_u$ is made independent from $u$ and, in our assumptions, equal to 1. With this we may finally set $C = (A_{rx}A_{ch}A_{tx})^{-1}$ to obtain

\[
X = b^o_{0} + \sum_{h=-\infty}^{\infty} y^o_{h} y^o_{h}(t) dt
\]

where we have highlighted the useful component $b^o_{0}$, the self-interference $I_u$ due to the other bits of the useful stream, the cross-interference $I_c$ due to the other transmitters, and the thermal noise contribution $I_n$.

Each of the non-useful components in the expression of $X$ may be assumed to be a zero-mean Gaussian random-variable. This is straightforward for the thermal-noise component but is also a sensible assumption for the self- and cross-interference that are weighted sums of the independent antipodal random variables $b^o_{h}$. In the light of this, the performance of the elementary decoder is uniquely determined by the variances $\sigma^2_x$, $\sigma^2_c$, and $\sigma^2_n$. In the following we will assume that the system is dominated by interference, i.e., that $\sigma^2_x + \sigma^2_c \gg \sigma^2_n$, so that the SIR

\[
\rho = \frac{1}{\sigma^2_x + \sigma^2_c}
\]

becomes the main performance figure.

The variances involved in the definition of SIR will be computed by averaging the terms $I_u$ and $I_c$ both on the transmitted bits ($E_y$) and on the possible spreading sequences ($E_y$), whose statistical features are dependent on the parameter $r$ obtaining

\[
\sigma^2_x = \sum_{h=-\infty}^{\infty} \sum_{h=\infty}^{\infty} E_y \left[(X^o_y)^2 \right],
\]

\[
\sigma^2_c = \sum_{u=1}^{U-1} \sum_{h=-\infty}^{\infty} \sum_{h=\infty}^{\infty} E_y \left[(X^o_y)^2 \right],
\]

where the last equality holds since it can be proven that, if we agree to use independent instances of the same sequence generator for each transmitter, then the expectation in $\sigma^2_c$ is independent of $u$ if $u \neq 0$. Hence, the cross-interference power can be factored in two terms, the first of which takes into account the ratios between the attenuation of the interfering component and the attenuation of the useful component, while the second takes into account the correlation of the waveforms corresponding to the unmodulated spreading sequences with different delays.
3. Performance Evaluation

In this section we determine, by numerical simulations, the optimum spreading and the corresponding maximum achievable SIR. The most commonly adopted shapes for UWB systems [6] are the Gaussian pulse \( \gamma(t) = \beta \exp(-t^2/(2\sigma^2)) \) and its second derivative \( \gamma(2)(t) \), where \( \beta \) is a constant used to adjust the amplitude of the pulse according to FCC norms. We consider a worst-case scenario in which the power of the interfering users at the receiver is equal to the power of the useful component, i.e. \( A_{ch} = A_{ch}^{0} \) for \( u = 1, \ldots , U - 1 \). We quantify the performance provided by chaos-based spreading policy over i.i.d. sequences as \( SU\% = 100(BR_o/BR_i-1) \) where \( BR_i \) and \( BR_o \) are the maximum values of BR we can achieve using i.i.d. and chaos-based sequences respectively.

3.1. Gaussian Pulse

In this case the shapes, at the input of the transmitter antenna and at the output of the receiver antenna, are represented by the Gaussian function and its 2-nd derivative respectively. As a reference example, we can consider a scenario with a spreading factor \( N = 127, U = U_{max} \) and a minimum value of link quality \( \rho_{min} = 9.54 \). The value of \( U_{max} \) is the maximum number of active users which allows us to obtain a value of \( \rho \geq \rho_{min} \) and is computed adopting i.i.d. sequences. This value is used to obtain the maximum Bit-Rate both for i.i.d. and chaos-based spreading policy. Figure 1 compares the SIR that is achieved using i.i.d. and chaos-based spreading respectively for different values of BR. The maximum achievable BR is evaluated intersecting the SIR curve with the horizontal line corresponding to \( \rho_{min} \).

In the case of i.i.d. sequences, we will indicate such a BR as \( BR_i \), whereas in the case of chaos-based spreading policy we have to use the dashed trend looking for the value of \( r \) which maximizes BR once that \( \rho \) is set equal to \( \rho_{min} \). We indicate this value as \( BR_o \) and, as one can see, we have \( BR_o \geq BR_i \).

Numerical simulations have been done for different spreading factors and active users and results are shown in Table 1. As one can see the BR we can achieve using chaos-based generation method is greater than the one obtained with i.i.d. sequences.

![Figure 1: SIR obtained using i.i.d. sequences vs optimum correlated sequences for a system with \( g(t) = \gamma(t) \), \( N = 127 \) and \( U = 14 \).](image1.png)

![Figure 2: Value of the parameter \( r \) which maximize \( \rho \) for a system with \( g(t) = \gamma(t) \), \( N = 127 \), \( U = 14 \) for different values of BR.](image2.png)

![Figure 3: BR as a function of the number of active users obtained using i.i.d. sequences and chaos-based sequences generation method for a system with \( g(t) = \gamma(t) \), \( N = 63 \) and \( \rho_{min} = 5.41 \).](image3.png)

Table 1: Performance in terms of BR using \( g(t) = \gamma(t) \) for different values of \( N \) and \( \rho_{min} \).

<table>
<thead>
<tr>
<th>( N )</th>
<th>( U_{max} )</th>
<th>( BR_i )</th>
<th>( BR_o )</th>
<th>( r_{opt} )</th>
<th>( SU% )</th>
<th>( \rho_{min} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>127</td>
<td>14</td>
<td>178.76</td>
<td>214.27</td>
<td>0.674</td>
<td>21.35</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>23</td>
<td>84.74</td>
<td>99.67</td>
<td>0.365</td>
<td>17.98</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>49</td>
<td>49.34</td>
<td>55.74</td>
<td>0.140</td>
<td>9.54</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>7</td>
<td>136.49</td>
<td>151.75</td>
<td>0.624</td>
<td>21.99</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>13</td>
<td>85.89</td>
<td>91.74</td>
<td>0.372</td>
<td>15.75</td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>27</td>
<td>135.09</td>
<td>145.18</td>
<td>0.952</td>
<td>19.75</td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>4</td>
<td>105.09</td>
<td>105.09</td>
<td>0.518</td>
<td>20.24</td>
<td></td>
</tr>
<tr>
<td>127</td>
<td>19</td>
<td>106.38</td>
<td>106.38</td>
<td>0.353</td>
<td>17.75</td>
<td></td>
</tr>
<tr>
<td>255</td>
<td>19</td>
<td>124.78</td>
<td>124.78</td>
<td>0.974</td>
<td>16.75</td>
<td></td>
</tr>
</tbody>
</table>

Note that this result is independent from the link quality \( \rho_{min} \) we have chosen. In Figure 2 we show the value of \( r_{opt} \) for different values of BR. A peculiarity is that this trend is not monotonous.
3.2. Second derivative of the Gaussian Pulse

We consider the case in which the transmitted pulse \( g(t) = \gamma^{(2)}(t) \). As in previous case, we can consider a typical scenario with \( N = 127 \) and \( \rho_{\text{min}} = 9.54 \). In this case we obtain \( U^{\text{max}} = 14 \). Figure 4 compares the SIR that is achieved using i.i.d. and chaos-based spreading respectively vs BR. When we use i.i.d. sequences we have \( BR_i = 76.34 \text{ Mbps} \) whereas when we adopt chaos-based spreading we have \( BR_o = 85.67 \text{ Mbps} \).

Figure 4: SIR obtained using i.i.d. sequences vs optimum correlated sequences for a system with \( g(t) = \gamma^{(2)}(t) \), \( U = 14 \) and \( N = 127 \).

In Figure 5 we show the value of \( r \) for different values of BR.

Figure 5: Value of the parameter \( r \) which maximize \( \rho \) for a system with \( g(t) = \gamma^{(2)}(t), N = 127, U = 14 \) for different values of BR.

In Table 2 we show the performance obtained with other spreading factors and link quality values. As one can see chaos-based spreading policy always outperform traditional i.i.d. sequences.

In Table 2 we show the performance obtained with other spreading factors and link quality values. As one can see chaos-based spreading policy always outperform traditional i.i.d. sequences.

In Table 2 we show the performance obtained with other spreading factors and link quality values. As one can see chaos-based spreading policy always outperform traditional i.i.d. sequences.

4. Conclusions

A chaos-based sequence generation method was proposed for synchronous DS-UWB WSNs. This method allows the reduction of MAI increasing the Bit-Rate. The proposed method was verified with traditional UWB pulse shapes and system parameters values. Numerical simulations has shown better performance of chaos-based spreading policy over traditional i.i.d. sequences.

References


Formulation of an assignment problem in wireless LANs

Hiroshi Tamura†, Masakazu Sengoku†† and Shoji Shinoda‡

†Niigata Institute of Technology
1719, Fujihashi, Kashiwazaki, Niigata 945-1195, Japan
††Faculty of Engineering, Niigata University
8050, Ikarashi-2nocho, Niigata 950-2181, Japan
‡Faculty of Science and Engineering, Chuo University
1-13-27, Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan

Email: tamura@iee.niit.ac.jp

Abstract— In wireless LAN, like IEEE802.11x, the number of accessible terminal is limited for each access point. Therefore, it is necessary to assign each terminal to the access points under the limitation. It is well-known that this problem can be solved in polynomial time by solving a maximum flow problem in graph theory.

However, there is a weak point in the above solution. If a terminal newly tries to access an access point, many terminals may change the access points. Therefore, this solution is not realistic. It is important to minimize the number of terminals that change the access points. In this paper, we propose a polynomial algorithm to solve the problem.

1. Introduction

In wireless LAN, like IEEE802.11x, the number of accessible terminal is limited for each access point. Therefore, it is necessary to assign each terminal to the access points under the limitation. For example, in Fig.1(a), the circle that includes an access point b is called the area of b. The area of b means that all terminals in the circle can communicate to b. Let the number of accessible terminals for each access point be three. In this case, we assign each terminal to the access points like Fig.1(b). In a simple case, we can apply the theory of flow networks in graphs to this problem. This problem can be solved in polynomial time by solving a maximum flow problem in graph theory[1]. First, we construct a flow network N as follows. The vertex set of N is \{b1,...,bh, v1,...,vk, s, t\} where bi represents an access point, vi represents a terminal, s is a source and t is a sink. A vertex pair (x,y) is an edge in N if and only if

i) x=s and y=bi or
ii) x=vi and y=t or
iii) x=bi, y=vj and vj is in the area of bi.

In case of i), the edge weight is the number of accessible terminals for bi, and in all other cases, an edge weight is one. Here, we calculate the maximum flow from s to t. In case of Fig.1(a), the flow network N is Fig.2(a). In Fig.2(a), we omit edge weights in case of 1. The maximum flow from s to t is Fig.2(b). Each bold edge means that the flow value of the edge is the maximum. We assign each terminal vj to the end access point bi of the bold edge being incident to vj (see Fig.1(b)).

However, there is a weak point in the above solution. If a terminal newly tries to access an access point, many terminals may change the access points. Therefore, this solution is not realistic. It is important to minimize the number of terminals that change the access points. In this paper, we propose a polynomial algorithm to solve the problem. We apply “Optimal Cardinality-Constrained Territory Map [2]” to this problem. The territory map is known as the Voronoi diagram[3] that is special kind of decomposition of an area in computational geometry. And
in this case, the number of elements in each territory is limited.

2. Formulation with graphs

We formulate this problem using graph theoretical terms as follows. We apply the formulation in Optimal Cardinality-Constrained Territory Map to this problem.

Let $G$ be a directed graph. Each vertex $v$ represents an access point. An directed edge $(u,v)$ means that there is a terminal $x$ in areas of $u$ and $v$, and $x$ communicates to the access point $u$ at this time.

Next, we set vertex weights of $G$. The weight of vertex $v$ represents the number of terminals that communicate to $v$, and $f(v)$ represents the maximum number of terminals that can communicate to $v$.

For example, in case of Fig.3 and $f(v)=5$ for each access point $v$, the directed graph is in Fig.4.

In this formulation, we obtain the following theorem.

[Theorem]

We assume that a terminal $x$ newly tries to communicate to an access point $A$ and $x$ is in the area of $A$. The terminal $x$ can communicate to $A$ if and only if in the constructed directed graph $G$, there is a directed path from $A$ to $B$, where the vertex weight of $B$ is less than $f(B)$.

In Fig.4, a terminal $x$ in area of $A$ is newly tries to communicate to $A$, there is a directed path from $A$ to $C$, where the vertex weight of $C$ is less than 5. Therefore, $x$ can communicate to $A$ in Fig.5 after changing access points of terminals.

Therefore, we should find paths from a vertex $v$ to vertices $u$, where the vertex weight of $u$ is less than $f(u)$. If the distance from $v$ to $u_0$ is the minimum, the distance means the minimum number of terminals that change the access points. Namely, we can obtain a solution by solving a single-source shortest-paths problem[4] in polynomial time.

Since the distance from $A$ to $C$ is 2 in Fig.4, the number of terminals that change the access points is 2 in Fig.5, where the underlined terminals changed the access points.

3. Formulation with hyper graphs

The formulation with graphs does not express all information in a wireless LAN about this assignment problem.
The directed graph corresponded to Fig.5 is Fig.6. Compared with Fig.4, the direction of edge (B,C) is converse, and there is a new edge between A and B in Fig.6. In the above case, the position of a2 turned into B from A in Fig.7. If the position of a1 instead of a2, turned into B from A, the directed graph is in Fig.8.

Fig.8 is different from Fig.6. Namely, this formulation does not express all information of this situation. Therefore, we have to change the access point of terminals using Fig.3 and Fig.4.

Here, we propose another formulation. We construct a directed hyper graph instead of a graph. If a terminal x is in areas of A, A1, A2,...,Ar and x communicates to A, we set a directed hyper edge from A to A1,...,Ar (see Fig.9). For example, in case of Fig.3, the hyper graph is Fig.10. Dotted lines represent a hyper edge.

The directed graph corresponded to Fig.5 is Fig.6. Compared with Fig.4, the direction of edge (B,C) is converse, and there is a new edge between A and B in Fig.6.

In the above case, the position of a2 turned into B from A in Fig.7. If the position of a1 instead of a2, turned into B from A, the directed graph is in Fig.8.

The hyper graph corresponded to Fig.6 is Fig.11.
The hyper graph corresponded to Fig.8 is Fig.12.

In this formulation, if a new terminal x is in areas A, A1, A2,…, Ar and x tries to communicate to A, we add a directed hyper edge from A1 to A2,…,Ar, to the hyper graph H, like Fig.9. If x is only in the area of A, we do not add an edge. Then we find a directed path from A to B in H, where the vertex weight is less than f(B), we reverse directions of all edges on the directed path and increase the vertex weight of A and B. Update of the hyper graph H is completed at this time. In this formulation, necessary operation is only on the hyper graph H. Namely, this formulation represents all information of a wireless LAN about this assignment problem.

4. Conclusion

In this paper, we formulate an assignment problem of wireless LANs using hyper graphs. This formulation represents all information of a wireless LAN about this assignment problem.

We will find assignment algorithms solving the problem and evaluate the algorithms with computer simulations. On practical use side, each terminal chooses access point distributedly. It is a future problem to take in these practical studies[5] for this modeling.

References

Slow Spiking and Variability of Inter-Spike Intervals near the Hopf Bifurcation in Simple Single Neuron Models

Junko Inoue†, Shinji Doi‡ and Sadatoshi Kumagai‡

†Faculty of Human Relation, Kyoto Koka Women’s University, Kyoto 615-0882 JAPAN
‡Graduate School of Engineering, Osaka University, Suita 565-0871 JAPAN

Email: jdoi@mail.koka.ac.jp, [doi, kumagai]@eei.eng.osaka-u.ac.jp

Abstract—We have shown that extraordinarily slow spontaneous spiking can appear near the Hopf bifurcation point in various (singularly-perturbed) single neuron models under a certain generic condition. This phenomenon raises a serious question about the classical categorization of neuronal firing (Class-I or Class-II neuron) by Hodgkin. This paper extends the result and examines the variability of inter-spike intervals with noise in detail and also discusses the relation to neuronal ability of information coding.

1. Introduction

Neurons are classified into two types: Class-I or Class-II neuron by the difference of the onset mechanism of repetitive discharge under the d.c. injection (homoclinic or Hopf bifurcation, resp.) [3, 4]. Class-I neurons possess a firing ability over a wide range of frequencies and thus possess a sufficient ability to encode information using the frequency or inter-spike intervals (ISIs). According to our recent finding, the Class-II neurons behave as Class-I neurons under a generic condition, and this ‘chimera’ neuron with both Class-I and II characteristics can produce very irregular (chaotic) spiking.

On the other hand, cortical neurons in vivo are known to fire very irregularly in response to a constant sensory stimulus and there is a growing interest in this highly variable (irregular) ISIs [6, 7]. In this paper, we study the noise effects on the spiking of the (slightly modified) Hodgkin-Huxley neuron model which has both the Class-I and II properties, and examine the spiking variability in detail. In particular, we show that the slight modification of a time constant leads to an essentially different spiking and the variability of ISIs even in the presence of noise.

2. Single Neuron and Its Input-Output Relations

Hodgkin-Huxley Equations

In this paper we consider the following slightly modified Hodgkin-Huxley (HH) equations [2]:

\[
dv/dt = G(v, m, n, h) + I(t), \quad (1a)
\]

\[
dm/dt = (m^\infty(v) - m) / \tau_m, \quad (1b)
\]

\[
dn/dt = (n^\infty(v) - n) / \tau_n, \quad (1c)
\]

\[
dh/dt = (h^\infty(v) - h) / \tau_h, \quad (1d)
\]

\[
G(v, m, n, h) = I_{ion}(v, m, h) + I_E(v, n) + I_I(v),
\]

where \(v\) is the membrane potential in mV, \(m, n, h\) are the dimensionless gating variables. \(I(t)\) denotes an input current (\(\mu\)A/cm\(^2\)) externally applied to a neuron.

The only modification from the original HH equations [5] is the introduction of the (dimensionless) ‘time constants’ \(\tau_m, \tau_n, \tau_h\); in the case of \(\tau_m = \tau_n = \tau_h = 1\), these equations become the original HH equations. It is shown that a slight modification of \(\tau_h\) changes the HH dynamics to a completely different one, with chaotic spiking and very long ISIs appearing in a generic manner, although the initiation mechanism of repetitive firing is a simple Hopf bifurcation [1].

Random Inputs to Neuron

Neurons receive many excitatory and inhibitory synaptic inputs from other neurons and these inputs have apparently random nature. The depolarization \(V(t)\) of the corresponding conductance-based or HH-type model such as the modified HH equations (1) is described by a random differential equation:

\[
dV = G(v, m, n, h) dt + \sum_{k=1}^{m} e_k dP^k_E(t) + \sum_{l=1}^{n} i_l dP^l_I(t),
\]

where \(e_k\) and \(i_l\) are the amplitudes of excitatory and inhibitory postsynaptic inputs, respectively. \(P^k_E\) and \(P^l_I\) are independent Poisson processes with mean rate \(\lambda^k_E\) and \(\lambda^l_I\).

Under the condition that all \(\lambda^k_E\) and \(\lambda^l_I\) diverge while simultaneously all postsynaptic inputs \((e_k, i_l)\) become
infinite small, the depolarized potential $V(t)$ is approximately described by

$$dV/dt = G(v, m, n, h) + I_{\text{ext}} + \sigma \xi(t),$$

(2)

where the constant bias term $I_{\text{ext}}$ is the mean input to a neuron, $\xi(t)$ is the Gaussian white noise, $\sigma$ is the noise intensity and the term $I_{\text{ext}} + \sigma \xi(t)$ denotes the total input current to a neuron. Thus, we examine the noise effects on the input-output relation of both the original HH equations and the modified HH equations ($\tau_f = 2$), where we set $I(t) \equiv I_{\text{ext}} + \sigma \xi(t)$ in eq. (1a).

In the following, we consider the coefficient of variation (CV) of ISIs which is often used as a measure of spike train irregularity. The CV (dimensionless) of a random variable $T$ is defined using the expectation (mean) and the variance of $T$ as $CV \equiv \sqrt{\text{Var}(T)/E(T)}$.

3. Spiking and Input-Output Relation of the Original HH Equations in the Presence of Noise

Figure 1(a) and (b) illustrate the period of the periodic firing of the original HH equations as a function of the external current $I_{\text{ext}}$ (a constant bias of synaptic inputs) when noise is absent: $\sigma = 0$. We can see that the period of stable periodic solutions (filled circles) varies in a range from several milliseconds to twenty milliseconds.

In Fig.1(c)-(f), the mean and CV of ISIs in the presence of noise are plotted as a function of $I_{\text{ext}}$ near the onset point of repetitive spiking: HB1 (Hopf bifurcation). Figure 1(c) shows that very small noise ($\sigma = 0.1$) does not change the behavior of the original HH equations and firing is very regular (very low CV). When $\sigma = 0.5$ (Fig.1(d)), both the mean ISI and CV drastically increase for rather small $I_{\text{ext}}$ (in the range between DC3 and DC1 of panel (b)). This reason is, however, very simple and not so interesting; in this range of $I_{\text{ext}}$, a stable equilibrium point and periodic orbit coexist (multi-stability). Because the noise intensity is not so high, a solution is trapped near either the stable equilibrium or the stable limit cycle and stays there for a while, and escapes from the attractors ‘by chance.’ Thus the mean and CV of ISIs increase significantly. The return maps of ISIs (the panel of $I_{\text{ext}}$ vs. $I_{\text{ext}}$) and the waveforms $v(t)$ at $I_{\text{ext}} = 7.0, 8.0$ clearly show this phenomenon. As $I_{\text{ext}}$ increases, the time the trajectory or orbit of the HH equa-
Figure 2: Noise effects on spiking and ISI variability in the HH neuron when the time constant is slightly changed: $\bar{\tau}_h = 2$. 

\[ I_{ext} = 9.57 \]
tions spends near the stable equilibrium becomes shorter since the stability of the stable equilibrium point becomes weaker as $I_{\text{ext}}$ increases and approaches the Hopf bifurcation point HB1. For large $I_{\text{ext}}=9.0$, such trapping near the equilibrium is suppressed and firing becomes very periodic. Note that there is still a multi-stability of stable equilibrium and limit cycle at this value of $I_{\text{ext}}$ (see the panel (b)).

When $\sigma = 1$, the transition between the stable equilibrium and the periodic orbit easily occurs. The variation range of the mean ISI is small compared to the small noise case, but the CV takes values between zero and unity linearly depending on $I_{\text{ext}}$. In the case of much larger noise: $\sigma = 5$, the firing is very regular and does not depend on $I_{\text{ext}}$.

4. Slight Increase of a Time Constant Changes the Spiking in the HH Neuron Much

Figure 2(a) and (b) illustrate the period of the periodic solutions of the modified HH equations in the noise-free case ($\sigma = 0$), where the time constant $\tau_h$ is just doubled. The only difference between the $\tau_h = 2$ case and the original one is the relative position of the double-cycle (DC) bifurcation point DC3; DC3 and thus DC2 points in Fig.2(b) are on the ‘right’ of the Hopf bifurcation point HB1 while they are on the ‘left’ of HB1 in Fig.1(b). As a result, there is no stable solution (stable equilibrium or simple firing) seen between HB1 and DC3 (only ‘simple solutions’ are shown in this bifurcation diagram obtained by AUTO). This slight change of the relative position of the DC bifurcation points, however, produces essentially different behavior: complicated periodic and chaotic firing with very long ISIs appear in the left of DC3 [1].

Figure 2(c) shows the mean and CV of ISIs as functions of $I_{\text{ext}}$ near the Hopf bifurcation point HB1 in (b). In contrast with the original HH equations, we can see very complicated structures in CV even in the absence of noise. We can also see the significant increase of the mean ISI near the Hopf bifurcation point HB1, although the time constant $\tau_h$ is just doubled. As is shown in (d), the small noise ($\sigma = 0.01$) does not destroy this structure. The ISI return map ($I_{\text{th}+1}$ vs. $I_\text{th}$) presents a finite lattice structure. This means that the spiking is essentially different from that of the original HH equations and that it is the consequence of the inherent nonlinear dynamics induced by the slight change of $\tau_h$.

As the noise intensity increases, much longer ISIs and high CVs appear near HB1 (the $\sigma = 0.04$ case of Fig.2(c)). The effect of noise is not uniform and much pronounced at smaller $I_{\text{ext}}$. The ISI return maps are completely different from that of the original HH equations also in this case.

When $\sigma = 0.1$, the fine structures is smoothed by noise, but the long ISIs still appear. Note that no noise-induced phenomenon appears in the original HH equations at this small noise intensity (Fig.1(c)); the increase of $\tau_h$ also changes the noise sensitivity of the HH neuron much. Further increase of noise ($\sigma = 1$) eventually induces spiking with short ISIs and moderate CV irrespectively of the constant bias $I_{\text{ext}}$.

5. Discussion

We have examined the spiking and its variability in the modified HH equations in the presence of noise. Slight increase of $\tau_h$ completely changes the spiking (even in the noise-free case). Noise increased the mean ISI and ISI variability (CV) also in the original HH equations. This phenomenon is, however, neither interesting nor unexpected; the mechanism is very simple (‘by-chance’ mechanism explained in Sect.3).

The modified HH equations have much richer dynamics; relatively small noise produces spiking with moderate and high CVs over a wide range of $I_{\text{ext}}$. The effect of noise is not just the blurring of the deterministic structure; noise-enhanced phenomena could appear. It is worth noting that the slight increase of $\tau_h$ also increased the noise-sensitivity. We consider that these results shown in the present paper are important in the study of neuronal coding, especially in the study of spiking variability of cortical neurons.

Note that all phenomena presented here are generic ones and can appear in the case of $\tau_h$ increase other than $\tau_{III}$, and in other simplified neuronal models such as the FitzHugh-Nagumo model also [1].

Acknowledgments

This research was partially supported by the Japanese Ministry of Education, Science, Sports and Culture, Grant-in-Aid (#18500172) and by Artificial Intelligence Research Promotion Foundation.

References

Theta Phase Coding in Entorhinal-Hippocampal Loop Circuits Organized by Afferent Signals

Jun Igarashi, Hatsuo Hayashi and Katumi Tateno

Department of Brain Science and Engineering,
Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology
2-4 Hibikino, Wakamatsu-ku, Kitakyushu 808-0196, Japan
Email: igarashi-jun@edu.brain.kyutech.ac.jp, hayashi@brain.kyutech.ac.jp, tateno@brain.kyutech.ac.jp

Abstract—We will propose a possible model of the entorhinal cortex layer II in which afferent signals are encoded by firing phases with respect to a theta rhythm. Although the entorhinal cortex layer II (ECII) sends signals to the hippocampus and the hippocampus sends signals back to the entorhinal cortex layer V (ECV), projections from ECV to ECII also exist in the entorhinal cortex. Moreover, it has been suggested that the ECII is also a theta rhythm generator. These imply that enhanced projections from ECV to ECII form entorhinal-hippocampal loop circuits and reverberation activity superimposed on a theta rhythm occurs in the loop circuits. First, we will briefly review our previous work on a network model of ECII that reproduces a theta rhythm. Secondly, we will show that loop connections from stellate cells to other ones are selectively potentiated in ECII by afferent signals. As a result, stellate cells fire successively within each theta cycle through enhanced loop connections. Our model also reproduces the phase precession of stellate cell firing within a theta cycle.

1. Introduction

It has been supposed that a sequence of places is encoded in the hippocampus by phases of the place cell firing with respect to a theta rhythm [1, 2], and the phase of the place cell firing advances within a theta cycle when the animal traverses a place field [3]. Recently, it has also been suggested that the entorhinal cortex plays important roles in place recognition and phase coding [4, 5]. Moreover, neurons in the medial entorhinal cortex layer II are theta-modulated and show phase precession [6].

The entorhinal cortex is located between the neocortex and the hippocampus. Layer II of the EC (ECII) sends signals to the hippocampus and layer V of the EC (ECV) receives signals from the hippocampus. In addition to this major pathway, projection from ECV to ECII also exists in the entorhinal cortex [7-9]. Thus, as the frequency of the activity in afferent pathways to the entorhinal cortex is increased, reverberation of activity consequently occurs in the entorhinal-hippocampal loop circuits [10, 11]. It has also been reported that long-term potentiation (LTP) is induced in deep-to-superficial layer projections in the entorhinal cortex by stimulation of the deep layer [12]. This implies that reverberation readily occurs in enhanced entorhinal-hippocampal loop circuits. On the other hand, although theta rhythm is observed in a wide range of the hippocampus [13], it has been suggested that EC is also a theta rhythm generator [14, 15] and that the theta rhythm results from synchronization of subthreshold oscillations in stellate cells [16].

When two afferent signals that have different frequencies are fed to different stellate cells respectively, firing phase of those stellate cells would be different because of different rise times of integrated EPSPs. If the time difference between the firings is 20-30 ms, the time difference between the reentrance signal through a loop circuit from a stellate cell and the spike of another stellate cell would be within a LTP time window because it takes signals about 20 ms to go around the loop circuits.

In this paper, we will propose a possible model of the entorhinal cortex layer II in which afferent signals are encoded by firing phases with respect to a theta rhythm. First, we will briefly review our previous work on a network model of ECII that reproduces a theta rhythm. Secondly, we will show that loop connections from stellate cells to other ones are selectively potentiated in ECII by afferent signals using a network model of ECII with loop connections that mimic entorhinal-hippocampal loop circuits. As a result, stellate cells fire successively within each theta cycle like a theta phase coding of places. Our model also reproduces the phase precession of stellate cell firing within a theta cycle.

2. Methods

2.1. Network Model of the Entorhinal Cortex Layer II

The present network model consists of 30 stellate cells (open circles) and one inhibitory interneuron (filled circle), as shown in Fig. 1. The stellate cells are not connected to each other through excitatory synapses, while the stellate cells are mutually inhibited via the interneuron. Periodic signals are fed to some stellate cells through excitatory synapses. All of the stellate cells receive excitatory random synaptic input independently because EPSPs occur spontaneously at a low rate in stellate cells in ECII [17]. The distribution of interspike intervals of the random input is Gaussian; the average interval is 1 s and the standard deviation is 250 ms.
The present network model of ECII contains loop connections that mimic the entorhinal-hippocampal loop circuits. The stellate cells are connected through loop connections in an all-to-all fashion, and the transmission delay through the loop connections is 20 ms. The loop connections are modified through a Mexican hat type learning rule as shown in Fig. 3b.

Fig. 1 Network model of the entorhinal cortex layer II with loop connections that mimic entorhinal-hippocampal loop circuits.

2.2. Models of Entorhinal Cortex Layer II Neurons

Multicompartmental model of the stellate cell developed by Fransén [18, 19] was used, though a few modifications were made. Principal modifications are as follows. Constant bias current was injected to the soma compartment as a muscarinic ionic current, which has a large time constant, and voltage dependence of some rate constants were slightly shifted along the voltage axis, so that isolated single stellate cell causes subthreshold oscillation spontaneously at about 5.5 Hz. The stellate cell consequently fires at about 3.5 Hz at apexes of the subthreshold oscillation.

The multicompartmental model of the interneuron used here is exactly the same as that developed by Fransén [18].

3. Results

3.1. Synchronization of Subthreshold Oscillations in Stellate Cells

We will here briefly review our previous work on synchronization of subthreshold oscillations [20]. Figure 2 shows superimposed firing patterns of 30 stellate cells. There is no connection between the interneuron and the stellate cells in Fig. 2a. In this case, stellate cells receiving low-rate random inputs often fire and are not synchronized. On the other hand, if stellate cells are mutually inhibited via the interneuron, the stellate cells occasionally fire at about 0.73 Hz around apexes of subthreshold oscillations, and subthreshold oscillations are synchronized due to recurrent inhibition (Fig. 2b). Occasional firings of stellate cells due to low-rate random inputs activate the interneuron, and then stellate cells are inhibited in synchrony.

Fig. 2 Synchronization of subthreshold oscillations. (a) Without recurrent inhibition. Subthreshold oscillations are not synchronized. (b) Subthreshold oscillations are synchronized due to recurrent inhibition via the interneuron.

3.2. Time Difference between Spikes of Stellate Cells Elicited by Signals whose Frequencies are Different

When a periodic stimulus was fed to a stellate cell in the network where subthreshold oscillations were synchronized, firing phase of the stellate cell with respect to the subthreshold oscillations advanced with the stimulus frequency. This is because the rise time of integrated EPSP induced by higher-frequency stimulation is faster than that induced by lower-frequency stimulation.

Therefore, when two different signals were fed to two stellate cells in the network respectively, the difference between firing phases of the stellate cells depended on the difference between stimulus frequencies. The spike time difference increased with increase in the difference between frequencies of stimuli that were fed to the stellate cells.

3.3. Synaptic Potentiation of Loop Connections by Afferent Signals to Stellate Cells

Histogram of the relative timing of spikes in the stellate cells is shown in Fig. 3a. These stellate cells are hereafter referred to as the stellate cells, A and B, so that the spike time difference $\Delta T = (\text{spike time of the stellate cell B}) - (\text{spike time of the stellate cell A})$. It should also be noted that the abscissa in Fig. 3a represents the difference between the spike time of the stellate cell A delayed by 20 ms and the spike time of the stellate cell B because the preceding spike of the cell A interacts with the following spike of the cell B after 20 ms through a loop connection.

The stellate cell A was always stimulated at 40 Hz, and the stellate cell B was stimulated at frequencies below 35 Hz. A Mexican hat type learning rule is shown in Fig. 3b, but enhancement of loop connections was not yet permitted here. The average spike timing shifted from negative to positive and was within the LTP time window of $\pm 10$ ms when the stellate cell B was stimulated at
around 30 Hz. These results indicate that pairs of afferent signals whose frequencies are appropriately different can be coincidently selected by the Mexican hat learning rule.

**Fig. 3** Histogram of the relative timing of spikes in the stellate cells (a) and a Mexican hat type learning rule (b). The stellate cell A was always stimulated at 40 Hz. Figures in the panel (a) indicate frequencies of stimulation of the stellate cell B.

Fig. 4 shows enhancement of the loop connection through the Mexican hat learning rule by pairs of afferent signals. The stellate cell A was always stimulated at 40 Hz. Figures in the panel show frequencies of stimulation of the stellate cell B. The synaptic conductance of the loop connection was enhanced when the stellate cell B was stimulated at around 30 Hz, while the synaptic conductance was not enhanced when the stellate cell B was stimulated at 37.5 or below 20 Hz. This clearly shows selective synaptic enhancement by pairs of afferent signals.

**3.4. Sequence Firing of Stellate Cells**

Four stellate cells A-D were stimulated simultaneously at 40, 30, 20 and 10 Hz respectively for the first period of 40 s in order to potentiate loop connections; consequently the loop connection from the cell A to the cell B was potentiated. Then, stellate cells B-E were stimulated simultaneously at 40, 30, 20 and 10 Hz respectively for the next period of 40 s and the loop connection from the cell B to the cell C was potentiated. After that, stellate cells, C-F, D-G and E-H, were stimulated every 40 s in the same way. Loop connections, C->D, D->E and E->F, were successively potentiated.

After the establishment of potentiation of selected loop connections, the stellate cells B-D fired successively after firing of the stellate cell A when the stellate cell A alone was stimulate at 40 Hz. Figure 5 shows an example of the firing patterns of the stellate cells A-D. Intervals between successive spikes are 25-55 ms.

If the stellate cells were successively stimulated at 40 Hz in the order of A to G, the spike of each stellate cell advances within a cycle of subthreshold oscillations (theta rhythm).

**Fig. 5** Firing patterns of the stellate cell A-D. Potentiation of selected loop connections is established. The stellate cell A alone was stimulated at 40 Hz. The stellate cells A-D fired successively.

**4. Discussion**

The present network model showed synchronous subthreshold oscillations in stellate cells as reproduction of a theta rhythm. Stellate cell firing was paced by the subthreshold oscillations and its phase with respect to the synchronous subthreshold oscillations depended on the stimulus frequency. Pairs of afferent signals whose frequencies were properly different were therefore selected by virtue of a Mexican hat learning rule and loop connections that mimicked entorhinal-hippocampal loop circuits, and loop connections were selectively potentiated by those pairs of afferent signals. As a result, stellate cells fired successively within a cycle of subthreshold oscillations when the first stellate cell alone was stimulated.

In the present model, although physiological evidence for the learning rule in the deep-to-superficial layer projection in ECII has not been provided, a Mexican hat learning rule was adopted for coincidence detection of
reentrance signals through loop connections and other afferent signals to ECII. It would however be possible to detect coincidence of those signals by an asymmetric spike-timing-dependent learning rule that has been found in many brain areas. Only slight change in the difference between frequencies of pairs of afferent signals would be required in order to get into the LTP time window.

Recently, it has been suggested that spatial information is represented upstream of the hippocampus in the medial entorhinal cortex as mentioned above. Grid cells found in the entorhinal cortex are activated at vertexes of a grid of equilateral triangles [5]. Such grid cells in ECII are theta-modulated and show phase precession [6]. The present model of the entorhinal cortex with loop connections may transform afferent signals into metric representation, and stellate cells show phase precession. The present model might reflect a part of properties of the grid cell though we have not reproduced a grid-like firing structure. Properties and functions of the grid cells however are still remains of controversy.

Acknowledgments

The authors are grateful to Dr. M. Yoshida for fruitful suggestions. This work was partly supported by the Japan Ministry of Education, Culture, Sports, Science and Technology (The 21st Century Center of Excellence Program in Kyushu Institute of Technology, entitled “World of Brain Computing Interwoven out of Animals and Robots,” granted in 2003).

References

A Hierarchical Brain Inspired Computing System

Christopher Johansson† and Anders Lansner†

† School of Computer Science and Communication
Royal Institute of Technology and Stockholm University
100 44 Stockholm - Sweden
Email: cjo@nada.kth.se, ala@nada.kth.se

Abstract—In this paper we describe an abstract hierarchical model of the neocortex and its implementation in a connectionist network. This network has convergent feed-forward projections going from lower to higher levels in the processing hierarchy implementing competitive learning. There are also associative divergent backward projections going from higher to lower levels and associative recurrent projections at each level. Here we show that this type of multi-level attractor network extends the capabilities of a single layer network by enabling autoassociative storage of patterns not possible to store in the latter.

1. Introduction

We envision that in the coming decades, information processing systems based upon brain-like neural systems will become increasingly capable and start to replace or enhance current systems in tasks such as image processing, man-machine interfaces, speech recognition, robotic and, autonomous control systems. The development of this type of artificial neural systems is enabled and promoted by the progress in brain research and the rapidly increasing computer power available in cluster computers such as e.g. IBM’s Blue Gene/L (BG/L) and in molecular-scale computing. Ensuring error free operation in these kinds of massively parallel computers is becoming increasingly costly, but by using error redundant brain-like algorithms and architectures these cost could be kept low because errors in the computations and communication can be compensated for by the algorithms. Thus this type of algorithms holds the promise to be amenable for implementation in the next generation of molecular-scale electronics that is projected to be inherently noisy due to quantum effects [1, 2].

Here we evaluate the functionality of an abstract hierarchical network model of neocortex. This abstract model is based on the modularization of neocortex into hyper- and minicolumns and a hierarchy of modules corresponding to cortical areas. The minicolumns are the functional units and these are grouped into hypercolumns and several hypercolumns make up an area module. This model is implemented with neural networks that use local learning rules such as Hebbian learning. The role of the hypercolumns is to regulate the activity in the networks and they do this by implementing a soft winner-take-all function of the units’ activities.

A group of hypercolumns in a lower area forms the input to a hypercolumn in a higher order area and these groups are referred to as receptive fields. Self-organization by means of clustering and competitive learning is used to set up and train these forward projections. Divergent and sparse back-projections from the higher to lower order areas and recurrent projections in each area are setup and trained with a Hebbian learning rule. Here we only implement networks with a single low order input and one higher order hidden layer, but it is straightforward to generalize this architecture to a hierarchical multi-layered setup where the hidden layer provides input to a third layer and so on.

Important and characteristic features of this brain inspired information processing system are parallelism, redundancy, self-organization, and adaptivity. The modular design and local learning rules ensures that the system can be efficiently implemented on distributed computational hardware, e.g. a cluster.

The system that we discuss here can be seen as an attractor network with a preprocessing stage that decorrelates data. Such a system is similar to that studied by Bartlett and Sejnowski for invariant face recognition [3]. Similar to our system, theirs uses competitive learning to create a projection from an input to a hidden population. Also, as in our system, the hidden population implements an autoassociative memory, but there is no back-projection from the hidden to the input population in the Bartlett and Sejnowski network. The idea of adding a feature extracting preprocessor to an associative memory has also been explored by other authors [4, 5]. In some work, the feature extraction step has been simplified to a random projection from the input to the hidden population in order to decorrelate data [6].

2. An Abstract Hierarchical Model of Neocortex

Here, we discuss one possible mapping of our model onto the mammalian neocortex and the starting point is the columnar structure of neocortex. In the neocortex, about 100 neurons are grouped into a minicolumn and approximately 100 minicolumns form a hypercolumn [7]. Because the pyramidal cells in layer II/III and V/VI are tightly connected by excitatory synapses [8] the minicolumn can be seen as the smallest functional unit in cortex. The purpose of the hypercolumn may be to normalize the activity of the layer II/III and V/VI
pyramidal cells in the minicolumns and to facilitate the competitive learning among the afferents to layer IV neurons. This normalization is supported by a group of inhibitory basket cells that receive projections from all minicolumns within a hypercolumn and project via inhibitory synapses back onto these minicolumns. Further, the layer IV stellate cells project to the pyramidal cells in layer III, but there are no direct projections within a minicolumn from either layer II/III or V/VI back onto these neurons. This means that information is mainly transmitted from layer IV neurons to the rest of the excitatory neurons within a minicolumn. This circuitry makes it possible to separate bottom-up data from top-down predictions. Discrepancies between these two data streams may trigger attentional and learning mechanisms.

Next we describe the hierarchical organization of hypercolumns and areas in our conceptual model, shown in Figure 1. The smallest functional units are the minicolumns and all connections are formed between them. The building block of this hierarchical structure is a layer of recurrently interconnected hypercolumns corresponding to a cortical area, together forming an attractor network. The recurrent connectivity is mainly formed by connections between neurons in layers II/III and V/VI. Therefore these four layers are taken together into one functional auto-associative layer in our model.

The cortical areas of hypercolumns are connected by convergent forward projections and divergent backward projections. The forward projections originate on the presynaptic side from a local group of hypercolumns and converge to target a single or a few selected hypercolumns on the postsynaptic side. These projections originate in layer II/III on the presynaptic side and are received in layer IV on the postsynaptic side where minicolumns with specialized receptive fields are located [9]. The backward projections, originating mainly from layer V/VI and target neurons in layer II/III and V/VI of the receiving cortical area [9].

The lower areas drive the higher areas, which in turn modulate the activity in the lower areas (“top-down expectations”). In the neocortex the higher levels of the model corresponds to the frontal, parietal and temporal areas.

The forward projections are used to extract features from the preceding lower order area, or the input from LGN in the case of area V1, and in this model these connections do not need to keep information that enables a reconstruction of the input data. Instead, reconstruction of the input is taken care of by the back-projections. In this model we use competitive learning for feature extraction, which has been proven to be a simple and powerful method [10, 11]. It can also be aggregated and used in a hierarchical setup thus achieving accurate and invariant pattern recognition, e.g. with slow learning [12]. Based on arguments of computational efficiency and neural energy efficiency it can be argued that the neural code should be sparse and the advantages with this type of coding have been demonstrated in several papers [3-6].

From an algorithmic perspective, implementing competitive learning in many small sub-regions of the input, i.e. the receptive fields, instead of applying it to the entire input is a divide-and-conquer strategy. A major advantage of this strategy is that it makes parallel implementation trivial.

On a global scale, several the hierarchies shown in Figure 1 operate in parallel to process data from different sensory modalities. These different systems have long-range recurrent projections mainly via layers V/VI between them. Further, the highest order areas are likely to be shared between several such systems.

3. Implementation of the Cortical Model

In this section we present a computational implementation of a subsection taken from the conceptual hierarchical cortical model previously outlined. It includes only two interacting populations, but is straightforward to generalize to a hierarchical setup. The input population has forward projections, feeding into the hypercolumns of the hidden population. These projection are trained with a Hebbian projection and a feedback projection back to the input representation. The hidden population has forward projections, feeding into the hypercolumns of the hidden population. These projection are trained with competitive selective learning (CSL) [13]. CSL can theoretically achieve a partitioning that is optimal in the sense that it minimizes the variance of the input data representation. The hidden population has a recurrent projection and a feedback projection back to the input population. Both of these are trained with a Hebbian learning rule, the BCPNN [14-16].

The forward projections, that implement competitive learning, take their inputs from the receptive field in the input population as shown in Figure 2. The units in the
hypercolumns in the receptive field form an input vector $x^h$ that is used in the competitive learning. The units within each hypercolumn in the hidden population compete for activation and here we set the unit that has the smallest $L_2$-norm between its weight vector $w_i^h$ and the input $x^h$ active;

$$y_i^h = \begin{cases} 1 & \text{if } \min \sum (x^h - w_i^h)^2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Figure 2 shows an input population with three receptive fields. The receptive field $h=2$ in the center is composed of eight hypercolumns with two units in each and the corresponding input vector $x^h$ is composed of 16 elements. Every one of these units connect to each and one of the units in the hidden hypercolumn indexed with $h=2$ and for each unit $i$ in this hidden population a weight vector $w_i^h$ is updated by eq. (2).

$$w_i^h = \begin{cases} w_i^h + \frac{x^h - w_i^h}{\tau_C} & \text{if unit } i \text{ wins in hypercolumn } h \\ w_i^h & \text{otherwise} \end{cases} \quad (2)$$

4. Results

In the following experiments we have stored the 36 images, shown in Figure 3, in both a single layer attractor network and in the proposed two-layered network, and then tested how many of these that could be recalled without errors. As retrieval cues we used the stored images with different amounts of salt-and-pepper noise added. Figure 4 shows how the images are mapped onto the input population of the hierarchical system. The hierarchical network had 8 hypercolumns in the hidden population and these received input from 8 non-overlapping receptive fields that each covered 8x4 pixels of the input. The hidden population correspondingly consisted of hypercolumns, one for each receptive field, which had the same number of units (feature detectors).

Single layer attractor networks trained with a local Hebbian learning rule can perform impressively on tasks such as noise reduction, pattern completion, and prototype extraction given that the data is encoded in a suitable manner i.e. sparsely. But when densely coded and correlated patterns such as those in Figure 3 are used, single layer attractor networks perform poorly. Here we constructed a single layer network from the input population by supplying it with a recurrent projection. In the experiments with this network we found that it was not capable of retrieving any of the 36 stored images.
5. Discussion

In this paper we have presented an integrated memory system that combines an attractor neural network with a decorrelating and sparsifying preprocessing stage. This memory system can work with correlated input as opposed to simpler autoassociative memories based on single layer networks. Here we did not implement self-organized receptive fields, but this has been done previously with good results [17]. Setting up the receptive field in an informed way so that correlated hypercolumns in the input population are grouped into the same receptive field enables the competitive learning to find better and more compact representations. We have implemented very large networks with this architecture to run in parallel on cluster computers and also studied what could be gained by e.g. FPGA implementation [18]. Because these networks implement intrinsically local computations and are trained with local learning rules they parallelize very well which is important for both cluster computer and dedicated hardware implementations.

The proposed implementation of our model could be regarded as an instantiation of the hierarchical Bayesian inference machinery proposed by Lee and Mumford [19] as a model of the visual system. In their framework the computations does not proceed in a bottom-up serial fashion but occur concurrently in a number of loops throughout the entire hierarchy.

Investigating this kind of brain-like systems is interesting both from a neuroscience and engineering perspective. This type of model can enable the understanding of neural processing principles used in the brain and also function as a test bed for hypotheses about global scale brain function. Furthermore, the brain is a system that readily solves a number of problems that today’s computing systems at best solves poorly. Many researchers now believe that we are approaching an inflection point where the computational hardware used for running neural networks has become powerful enough to implement brain-like computing systems of a size approaching that of their biological equivalents and that this in turn will open up for numerous new applications.

References

Optimization at work in the cortex: Limit cycle computation, minimization of the total wiring length, and architectural computational classes

R. Stoop, St. Martignoli, and C. Wagner

Institute of Neuroinformatics ETHZ/UB, Winterthurerstr. 190, 8057 Zurich, Switzerland
Email: ruedi@ini.phys.ethz.ch

Abstract—Biological neocortical neurons are arranged in a columnar clustered architecture and coupled according to a bi-power law connection probability function. We discuss a computational framework in which recurrent connectivity is of obvious advantage. Taking columns as the basic computational elements, we use a network framework that allows us to scan a wide range of network types. We find that the information propagation and synchronizability saturate most quickly at high values for a close-to-biology model of the cortex, expressing the optimality of this network type. We finally come to the conclusion that the neocortical network optimizes information propagation and synchronizability under the constraint of constant total connection length, and that recurrent / nonrecurrent connectivities link to P / EXP computational classes.

1. Cortical columns: A local computation paradigm

The classical machine model of computation consists of digitally specified states, and specified operators that transform states one to another. For neuronal systems, it is, however, difficult to grasp what the underlying natural codes are, and how they are realized in the neuronal architecture. Answering the latter question is of great interest, as the information processing by means of biological neurons has efficiency and stability features that are unmatched by classical computation.

In the mammalian cortex, substructures can be identified that are devoted to specific sensory processing. Major areas are, for example, occupied by the auditory or by the visual cortex. Although serving different purposes, they follow quite closely a common building principle. In mammals, the primary visual pathway relays information from the retina across the lateral geniculate nucleus (LGN) to the striate cortex. Although the precise function of the LGN is unknown, it is suspected that LGN implements coding efficiency by canceling out redundant information. Striate cortex cells are predominantly orientation selective: They respond best to lines aligned with the main axis of their receptive field. For perpendicular stimuli, no or poor response is observed. The morphological study of this cortex exhibits vertical and horizontal structures. The vertical structures divide the cortex into six main layers (labeled I-VI from periphery to inside). In the horizontal direction, functional columns can be distinguished in which the neuronal density clusters. A more detailed analysis reveals that columns are composed of 50 – 100 minicolumns that comprise 80 – 100 neurons each except in the primate striate cortex, where the number is more than doubled. It appears that the basic unit of cortical operation is the minicolumn which measures of the order of 40 – 50 µm in transverse diameter, separated from adjacent minicolumns by vertical, cell-sparse zones. The minicolumn is produced by the iterative division of a small number of progenitor cells in the neuroepithelium. Physiologically observed functional columns are formed by the binding together of many minicolumns by common input and short range horizontal connections, where the number of minicolumns per column varies between 50 and 80. Their diameter varies between 300 and 500 µm. Long range intracortical projections link columns with similar functional properties. The best known example are orientation columns in the striate cortex. They contain 80 x 100 = 8000 neurons [1]. It appears that columnar identity is maintained in both forward and backward projections. The majority of the excitatory cells are pyramidal neurons that make up as much as 70 % of all neurons. They often spread across all layers, seemingly relating or summ ing up what goes on in layers II-III (which mostly relate to other cortical areas) or V-VI which relate more to functions and signals from and to within the body, such as, e.g., the motor cortex. Layer IV is the input layer. It can be seen as the amplification machine for the incoming signal from LGN, which then is relayed to layers III and II and from there to VI and V. Layer IV also provides a network of inhibitory neurons that shuts down the amplification process upon the corresponding signaling from the pyramidal cells mostly of layer V. The (essentially 5-) layered construction of the cortex can be interpreted as follows. We have basically one outside/inside neural network that each has several layers, possibly in order to overcome the linear separability constraint of neural networks. In support for achieving nonlinear separation boundaries, the network is helped by inhibitory neurons, whose task is not only to provide the shut-down mechanism in layer IV, but can be seen to contribute one additional layer (the Werbos construction). By means of their stronger (than excitatory) inhibitory synapses, they are also able to induce temporally coherent signals.
An extended visual object elicits neural activity across many different cortical columns, so that the object is encoded at different cortical locations. The lateral connections are thus needed for obtaining a coherent representation of a spatially extended input object. Roerig et al. [2] determined the distribution of lateral connections in the striate cortex of ferrets. They found that the connection probability for two neurons is not Gaussian, but has a power-law long tail. Intralayer connections are sufficiently dense so that active modules a little distance apart can become associatively linked. The network thus becomes a network of networks where the basic computing units are not neurons but (mini)columns. A natural question is in what sense this architectural setting could be computationally advantageous. The answer cannot be that short local connections are cheaper than long ones. Rather, it relates to the observation that on any level of neuronal interaction, information is broken down to noise. On this level, "events" are not only difficult to describe, but also their functional relevance is unclear.

Dynamical equations of biological neurons are approximations, whose qualities are often difficult to access. What is easily verified is that the biophysically best motivated neuron models, the Hodgkin-Huxley (H-H) and the Morris-Lecar equations, yield limit cycle solutions in the relevant regimes (see also Fig. 1). An explicit proof of this property for cortical pyramidal neurons is by the experimental evidence that coupled biological neurons lock along Arnold tongues, a phenomenon unique to limit-cycles. In the corresponding experiment, regular pulses were put on the axon of a neuron that was connected to a target pyramidal neuron. The latter was also driven by a current at, however, differing frequency. When sweeping over the product space of the frequencies, we observed the typical locking along Arnold tongues [3]. The same procedure was applied to simulations of detailed neuron structures (by NEURON), revealing that also in this case the coupled neurons quickly lock, in frequency and in phase.

Figure 1: Locking among H-H neurons $n_1$ and $n_2$, driven by unequal currents, bidirectionally excitatorily coupled over $\alpha$-type synapses of equal strength. Shown are voltage ($V$) traces and interspike ($I$) histograms of the two neurons $n_1$, $n_2$, respectively. A clear pattern in the generated firing events emerges, showing periods 2 and 3.

Figure 2: Arnold tongues of one-sided inhibitory coupling. $\Omega$ is the frequency ratio between the two intrinsic firing frequencies, and $K$ the strength of the synaptic coupling. The numbers indicate the order of the generated periodicities, $s$ the direction of the sweep of the inset.

Fig. 2 was generated from an electrophysiologically derived circle map [3]. The inset demonstrates that when the frequency ratio $\Omega = \frac{f_1}{f_2}$ changes along the line $s$, this is documented in the experimentally measured firing patterns, similarly to a log-book. The definition of $\Omega$ implies the invariance of the periodicity for firing frequencies multiplied by a common activity factor. In this way, the emerging firing pattern can be seen as the result of the computation from two analog driving current inputs $I_1, I_2$ [3]. This indicates that locking could provide a valuable means for information coding, in the frequency as well as in the temporal coding sense. The only necessary conditions for locking are 1) a sufficient separation between the modulations of the driving and the neuron’s intrinsic firing frequencies, and 2) a separation into weak (CLT-like) noisy drivings and strong or coordinated pulses among the locked neurons. Is such a situation likely to occur, taking into account the physiological facts? When a neuron is stimulated according to some regular pattern, this may trigger long-term (LTP) and short-term (STP) potentiation mechanisms, which may facilitate the synapses’ efficacy by a factor of 1.5. I.e., locking is a self-enhancing process. In the Arnold picture, this effect can be associated with an increase of $K$ by an identical factor. Moreover, in hippocampus, two clearly distinguishable classes of synapses have been found, that differ in their release probability by a factor of about 6, which also should transfer into similar classes of efficacy. The percentage of efficient synapses is relatively small ($\sim 15$ vs. $85\%$), which seems compatible with the situation we envisage. If we have about 100
active synapses of small efficacy from estimated 50 – 100 neurons needed for the firing of the target neuron, already one strong synapse (or a small number of temporally synchronized strong synapses) could be sufficient for inducing locking.

The sweeping experiments suggest that the Arnold coding scheme might be sufficiently robust to be effective for transient signals as well. To corroborate this, we simulated H-H neurons that we coupled reciprocally excitatorily / inhibitory via α-type synapses. For the experiments, we chose constant currents that would according to the Arnold tongues, produce e.g. periodicity 2. We then modulated these currents \( I(t), i = 1, 2 \) by a common driving function \( d(t) \) to \( I_i(t) := d(t) I_i(t) \). Examples of driving currents used are shown in Fig. 3. If the driving is sufficiently variable or sufficiently slowly varying, this is still equivalent to a CLT driving and the locking is preserved. Locking only fails in an intermediate range when the modulation substantially interferes with the neuron’s own firing frequency. In this case, ‘broad’ interspike interval distributions are generated. Because of the strong response and suppressed adaptation, this is the preferred experimental situation, although from the point of view of the working brain, it might be a rather exceptional one.

![Figure 3: Setting as in Fig. 1, but with variable modulations \( d(t) \) (first panel). Wildly varying currents (II) maintain locking if applied to both neurons. The same behavior is obtained for slowly modulated driving currents (III). Only when the driving current’s modulation is comparable to the neurons own preferred frequency (I), locking breaks down.](image)

These observations are important because a neuron, when on a limit-cycle solution, combines ideally noise and precise timing. The noise because of its CLT-close behavior, can be thought of as a driving source. Superimposed on noise, the neuron emits and receives precisely timed firing events. Provided that one agrees that precise timing has a meaning for the cortex, the exact relationship between the spikes emitted can be understood as a coding scheme. Information can be understood as a departure from background behavior, in a statistically significant way, which is achieved in the locking paradigm by setting off precise timing against noisy driving. By means of locking, the analogue information contained in two noisy drivings is converted into an essentially digital one, the firing periodicity, i.e., computation has been performed.

2. The cortical network: Minimizing wiring length

Having identified a putative origin of the cortical columnar structure, we investigated how on such structures the main cortical tasks that remain: information transfer and synchronizability, could be optimized. We implemented the biological situation as follows. Our basic units are the columns, which are characterized by a power-law decay outside a column [2]. The columns themselves behave rather independently, which suggests modeling them as variable, ideally: chaotic, units. This setting takes care of the recurrent biological architecture and of the fact that different columns can and often will, act independently, exchanging but small coupling. Although it is highly questionable whether a neuron in itself can be chaotic, when responding to local external driving, a modeling by chaotic network sites connected by diffusive coupling seems to make perfect sense.

We endowed the columns by a wiring framework. In a first step, we showed that single power-law fractally coupled (SPL) networks lead to small-world networks. This is already a strong indicator for the efficiency of the cortical structure. For intermediate power exponents, network characteristics (clustering coefficient \( C \) and mean path length \( L \)) emerge, the values \( L \approx 2.5 \) and \( C \approx 0.13 \) of which, obtained for \( N = 282 \) sites in the whole relevant range, are quite close to those of the neuronal network of the biological example \( C. elegans \) (\( L = 2.65, C = 0.28 \), for \( N = 282 \) [4]). Additional clustering structures and in particular bi-power law (BPL) networks that provide a more detailed modeling of the cortical network, leave these characteristics vastly unchanged. We then asked under what conditions a given information speed could be realized subject to the requirement of synchronizability. To this end, we derived the dependence of the information spread and the dependence of the synchronizability both on the mean number of neighbors. We went to the edge of synchronizability, from this we derived the information speed and compared in this way with concurrent networks. Although this number is substantially reduced in BPL compared to SPL networks, in purely random-coupled networks synchronizaton emerges for an even lower average number of connections per node (\( dg = 12 \)). Rather than this quantity, cortical neural networks appear to minimize the total connection length: Averaging over 10 networks, the simulations reveal a total connection length for the SPL network of 528876, and for the BPL network of 309306. Interestingly, with a
connection length of 390087, the random-coupled network places itself between the two paradigms.

3. Architecture: the generator of computational classes

To scrutinize the influence of architecture upon computation, we compared the computational results obtained for when the computational units are arranged in a feed-forward structure (characteristic for conventional artificial computation) vs. the arrangement in a recurrent structure (of which a bidirectional ring is the simplest manageable implementation). In order to reflect the local nature of computation, the building blocks for feed-forward and recurrent networks were chosen to be mono- and bi-directionally, respectively, coupled limit cycles. Whereas for two neurons the mono-directional interaction is described by the aforementioned circle-map, for bi-directional coupling, the case is more involved: If \( f(x_i) < 1 \), the corresponding map has the form \( f(x_i) = \frac{1}{2} \sqrt{\Omega^2 (g^i(x_i) - x_i)} - g^i(x_i) + x_i \), where \( \Omega = \frac{2}{3} \leq 1 \). If neuron 2 fires several times before neuron 1 does, the mapping is modified appropriately. In our numerical experiments, we initialized the neurons by intrinsic spiking frequencies drawn from the interval \([0.95, 1.05]\) and connected them with strengths from \([0.4, 0.8]\). After an initial transient, all neurons fired regularly with identical frequencies, but at individual phase shifts. An arriving signal was then implemented by a change of the intrinsic frequency of the first neuron, across the interval \([0.5, 1]\). This modified the responses of the neurons down the chain. In both architectures, from every neuron’s response (periodicity \( p \)) we could reconstruct the frequency of the input signal, up to a precision dictated by the width of the corresponding tongue and numerical resolution. It is customary to consider as the hallmark of computation information contraction. This amounts to measuring how fast the sensitivity towards the input signal decays as a function of the neighbor distance. To this end, we initialized all neurons with identical intrinsic frequencies, corresponding to an interspike interval \( T^i = 1 \) and chose coupling strengths of \( K^i = 0.6 \). Subsequently, the input to one (the ‘first’) neuron was changed to the maximally locking-resistant quasiperiodic golden mean signal \((T^j/T^i) \mod 1 = (\sqrt{5} - 1)/2\) and the generated signal bandwidth \( T_{\text{max}}^j - T_{\text{min}}^j \) was measured down the chain. Whereas in the feed-forward case a slow bandwidth decay described by a power law is observed, for the recurrent structure the decay is exponential.

The deviant behaviors have two important consequences. In the neocortex, feed-forward topologies are commonly associated with information transfer. In this context, the slowed-down destruction of information makes sense. In our framework of an objective measure of computation [5], the much faster destruction within the highly recurrent structures which are believed to be the predominant loci of the computation (master example: the cortical columns) makes sense as well. In this framework, the computation by a system can be measured by its ability to improve the predictability of a process [5], which is linked to a reduction of its prediction complexity [6]. As in our approach, the input signal to be predicted, one can argue that the computation per neuron is of superior quality in the case of the recurrent architecture, if compared to the feed-forward case. Such a property is often attributed to recurrent architectures, but usually lacks objective evidence. By mapping this finding to the time needed to arrive at the extinction of the signal down to a given precision (from where on the system’s behavior is exactly that of the unperturbed network and its prediction is trivial), we demonstrate that this is achieved much faster by the recurrent network (in an exponentially vs. a power-law decreasing demand in neurons and time). This, finally, provides the claimed (inverse) correlate to the fundamental computational hierarchy \( P \subset EXP \), within the architectures of neural networks.

References

The nontrivial grammar of Drosophila courtship language distinguishes among different genetic profiles

R. Stoop, C. Heid, and B. Arthur

Institute of Neuroinformatics ETHZ/UZH
Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: ruedi@ini.phys.ethz.ch

Abstract—The quest in behavioral neurogenetics is to understand how genes influence behavior. We provide an operational definition of behavior that allows us to quantify the similarity between different courtship behaviors of Drosophila, including that of the fru fruitless mutation. The method not only recovers all known salient phenomena of Drosophila courtship and makes them quantitative. We also obtain evidence that a variable, expressive body language underlies Drosophila courtship, which is particularly suited for the transmission of partner-specific information. In the presence of males carrying the fruitless gene, normal male Drosophila even perform a switch from male to female behavior.

1. Introduction

Courtship of the fruit fly Drosophila is a standard example of genetically hardwired behaviors. Whereas normal male Drosophila melanogaster court females only, using advanced genetic techniques applied to the fru gene [1] it is now possible to generate males that only court males or court both genders, and females that court males or only court females. This provides hard evidence that gene information in addition to how living beings are built also defines to a considerable extent how they behave [2]. Why animals court, is largely terra incognita. Since Drosophila is under strong evolutionary pressure, pre-copulation courtship could serve as an information platform for assessing a potential partner’s suitability for passing on genes. Here, we investigate how this purpose could be achieved within the genetical constraints. We find evidence of a nontrivial courtship language that provides the individual Drosophila with a large expression space. During the unrewarding courtship situation with a fruitless male, normal males even perform a contextual switch from male to female behavior. We anticipate our assay and the developed methodological approach to be a starting point for systematic investigations of the relationship between genes and behavior.

In Drosophila courtship experiments, single normal females in the immature, mature, and mated states were paired with single normal males in an observation chamber, which yielded 6 different groups of protagonists. To test genetical constraints, we also paired fruitless [1] mutant males with either mature females or mature normal males (4 more groups of protagonists). The emergent courtship behavior has tactile, gustatory, olfactory, acoustic and visual sensory dimensions. In pre-copulation courtship of D. melanogaster, visual information is salient and easy to access, which is why we focus on this component. This sensory reduction is not problematic, as one may faithfully reconstruct the multi-dimensional system from the single component [3]. Although our analysis is based on time series of one protagonist at a time, we also obtain insight into the interplay between partners.

We performed the visual recordings at the almost neuronal resolution of 30 frames per second. By basing our analysis on the 37 independent (“fundamental”) behavioral acts that, on this time-scale, compose the courtship behavior, we eliminated potential ambiguities in the definition of the courtship states. For the recordings, using frame-by-frame inspection and focusing on one single protagonist only, the starting points of the fundamental acts were detected. The corresponding indices were written down into files of symbols on which our further analysis is based. For each possible experimental constellation, five trials of the experiment were performed, using different individuals. The individuals were not previously screened for how efficiently they would court.

To estimate the courtship information content of the files, we evaluated the Shannon entropies \( H_i = -\sum p_i \log p_i \). In the formula, \( n_i \) is the number of symbols used, \( i \) indexes the fundamental acts (or their corresponding symbols) and \( p_i \) their probability of occurrence [3]. For similar behavior, we expect to measure similar values of \( H_i \). The averaged results obtained support all basic observations known to hold for Drosophila courtship: Overall, male Drosophila use a richer repertoire. Under suboptimal conditions, i.e., in the immature state (experiment 2) or when paired with fruitless males (experiment 5), the female repertoire is reduced. Males use a richer repertoire to court an already mated female (experiment 1 vs. experiment 4). In experiment 5 (mature female, fruitless male), a decreased activity is accompanied by a reduced repertoire by both protagonists. When normal males are paired with fruitless mutant males (experiment 3), the entropies support the visual observation that under these conditions the normal males adopt a female-like role and that the fruitless males express a very
pure male-like role. Otherwise, a crossing of the entropies would be obtained.

Taken as a measure of behavior, the Shannon entropy has profound deficits. $h_s$ depends strongly on the number of the symbols used, which may be partially responsible for the overall difference between female and male behavior (more symbols are accessible to males than to females). As a consequence, an unbiased comparison of information is restricted to within the female group (i.e. immature, mature, mated), or the male group (i.e. normal and fruitless males). Even more importantly, the Shannon entropy ignores the individuality of the symbols and the order in which they appear, which both are obvious important characteristics of behavior. Entirely different temporal sequences, composed of different dominant symbols, may therefore lead to identical values of $h_s$. The fact that $h_s$ is maximal for an equi-distributed random process (that would correspond to a fortuitous testing of all available symbols), is a further indicator of the difficulty of establishing a direct relationship between $h_s$ and behavior.

2. Scrutinizing the notion of behavior

If during pre-copulation courtship a potential partner’s suitability for passing on genes is assessed, we expect to observe a large behavioral variability within genetically normal animals and a reduced variability within the fruitless mutants. In addition, we expect that behavioral classes emerge that distinguish between normal and genetically mutant Drosophilae, between the different sexes, and possibly also between their different stages of sexual development.

For this analysis, the traditional tools courtship index (the fraction of the whole time spent by an animal in obvious courtship) and the mating success (the fraction of successful versus total attempts of copulation), are unsuited. They are both integral measures which cannot resolve more specific aspects of behavior. Precursors in our attempt to refine the description of courtship are behavioral transition graphs. One would expect to see genetic variations reflected in topological and metric modifications of these graphs, as well as in changed salient individual behaviors. Unfortunately, the usual Markov transition graph approach allows only few behavioral states to be considered. By concatenating several fine acts of behavior into macroscopic (i.e., temporally extended) states, we would lose the ability to address fine-scale issues. Moreover, first-order Markov models only take one transition step into account. As a consequence, no precise information about longer successions of fundamental acts, which seem essential for behavior, can be extracted. Higher-order models could be converted into first order, but only at the cost of exponential growth of the model in the order, which quickly leads beyond tractability. Moreover, comparability of the resulting graphs would be restricted to within each male/female class, because some of the fundamental acts only apply to females, others to males only.

To remedy this, we worked out an alternative way of characterizing behavior. Human courtship proceeds along stereotyped tracks: Take a girl for a coffee, go with her to the movies, walk her home (often several times), get invited for a last drink, etc. Here, we posit that for Drosophila a similar characterization applies. It is conceivable that precisely variable length sequences of fundamental acts provide a powerful description of behavior, similarly to the role of letters/words with respect to words in spoken language. Behavior is usually defined as “the aggregate of actions or reactions of an object or organism, usually in relation to the environment”. We propose to characterize behavior by its set of irreducible closed orbits of fundamental acts (indecomposable orbits of no further dissectable acts of behavior). As redundancy is an important element of biological information exchange, we may expect the characteristic behaviors to occur repeatedly, embedded within other activities, and assume that the characteristic behaviors are robust enough to statistically stand out from this random environment.

This definition of behavior leads directly to the following computational implementation: From the experimental data, we extract the set of irreducible closed orbits $s_i$, $i = 1 \ldots n$. A template vector of dimension $n$ is then formed by assigning to each of these orbits a particular vector component. As a function of their length $l$, the number of closed orbits can be expected to grow no more than $N(l) \sim e^{l h_{top}}$, where $h_{top}$ is the topological entropy of the process [3]. $l$ has to be chosen taking into account the size of the data and the behavioral context. In our case, we went up to a length of $l = 7$ symbols, for which we obtained 181 irreducible closed orbits. The template vector can be interpreted as defining the animal’s behavioral space with respect to all of the experiments considered. For each protagonist $j$, we constructed a behavioral vector $b_j$, by filling the entries of the template vector with the numbers of occurrences of the corresponding orbits.

Courtship behavior is not a continuous, but a highly variable process that occurs in bouts that can be seen as a prototype of complex behavior. The description of behavior in terms of closed sequences of fundamental acts is thus closely related to the dissection of complex chaotic motion into periodic orbits, which provides a variable-order Markov approximation to the underlying system. There is, however, more structure to it. Typically, courtship proceeds in a level-like organization, where only after completion of a certain level (possibly on the first attempt), one is allowed to access the next level.

In order to obtain a measure for behavioral similarity/dissimilarity, we use normalized behavioral vectors $b_i$, and the projectional properties of the scalar product. By projecting onto the orbits common to both compared behaviors, the scalar product assesses their similarity in a simple way. The influence of a multiple occurrence of orbits can be monitored by weighting the behavioral vector com-
ponentwise by an exponent $\beta \in \mathbb{R}$. This method is borrowed from statistical physics, where $\beta$’s role is that of an inverse temperature (e.g., [3]). In this way, we characterize the behavioral similarity between two protagonists by a one-parameter similarity family $M_{i,j} := (b_i, b_j)\beta$, $\beta \in \mathbb{R}$, where $\langle \cdot, \cdot \rangle$ denotes the scalar product of the behavioral vectors obtained for the protagonists $i$ and $j$, respectively. All pairwise similarities can be collected in a matrix $M(\beta) = (M_{i,j})$. For $\beta \geq 0$, large/small matrix entries indicate similar / dissimilar behaviors. For $\beta = 1$, the natural measure is obtained (i.e. the measure that is based on precisely how often cycles appear), whereas for $\beta = 0$, this information is reduced to the topological aspect (only the occurrence or absence of a particular cycle matters [3]). Bearing in mind that aberrant behavior of one single fly could completely dominate the average behavior, we will concentrate on the topological characterization. Note that our notion of similarity is only an approximate one that does not include transitivity.

Once the individual protagonists are listed according to their experimental group (see Table 1, for the correspondence ("\text{\textasciitilde}x")) between the matrix indices and the experiments), the similarities emerge that, as expected, exhibit a large individual variability among protagonists of the same class, but also appear to display common class behaviors.

| 1 = fruitless males-mature females | 2 = mature Females-fruitless males |
| 3 = fruitless males-normal males | 4 = normal Males-fruitless males |
| 5 = normal Males-mated females | 6 = mated Females-normal males |
| 7 = normal Males-mature females | 8 = mature Females-normal males |
| 9 = normal Males-immature females | 10 = immature Females-normal males |

Table 1: Correspondence between matrix indices and Drosophila behavioral vectors. Boldface letters identify the observed protagonists.

To assess the amount of bias introduced by a pooling from individual to class behavior, we performed two-sided mean-difference significance tests for the hypothesis that the sets of similarities obtained for two protagonists would originate from the same distribution. Although we expect this to underestimate the behavioral discriminability, in the result the original class structures are strongly reflected. Moreover, from the data, the identification of individuals sharing, or showing unusual, behavior within one class is possible. As an example, individual females – in particular mature females in the presence of normal males – obviously use only few behavioral alternatives. In contrast, the individuality of males in the presence of mature females appears to be important. This we take as an indication that it is the individual male behavioral response to the more stereotype

typical female behavioral patterns that defines the salient information pathways in Drosophila courtship.

At a given confidence level, majority voting can be used to estimate whether a class could distinguish between itself and another class. At a confidence level of $p=0.85$, all classes are distinct. At $p=0.95$, experiments involving the fruitless mutation, female behavior and male behavior can be distinguished. At confidence level $p=0.99$, the male behavior towards mature, and towards immature females can be distinguished from the rest. Above this confidence, only the male behavior towards mature females remains identifiable. Because these tests underestimate the behavioral discriminability, they not only confirm our approach’s ability to discriminate in a quantitative and fine-grained manner among the different behaviors involved in Drosophila courtship. We also conclude that pooling into classes of behaviors introduces only a small amount of a priori information. Pooling was performed by adding the behavioral vectors contained in a class and then renormalizing the obtained vector, giving rise to a matrix $\tilde{M}(\beta)_{i,j} := (\tilde{b}_i, \tilde{b}_j)\beta$, where $\tilde{b}_j, \tilde{b}_i, i, j \in [1, \ldots, 10]$, are the pooled behavioral class vectors. A density plot of $\tilde{M}(\beta = 0)$ as obtained from the experimental data is shown in Fig. 1a.

Figure 1: Class similarities based on a) the closed orbit analysis of the original data, b) the closed orbit analysis of surrogate data, c) on the symbol probabilities only (see text). Density plots over $[0, \text{maxcorr}] = [0, 1]$. Lighter shading indicates higher similarities, diagonal elements were set to 0. Red and blue circles indicate maximal and minimal similarities, respectively.

To corroborate that behavior is indeed expressed in terms of closed orbits, we shuffled the data files (surrogate data method) and compared the numbers of closed orbits from the original files with those from the surrogate files. We found that the surrogate data (which should not express behavior) contained substantially less closed orbits than the original files (see Fig. 2). Moreover, a substantial portion of the former (often as much as 50 percent) were not contained in any of the original files. This indicates that the closed orbits are not accidental, but rather a distinguished characteristic of behavior. Courtship behavior thus emerges similarly to the usage of language, where grammatical and syntactical rules govern the formation of the sentences.

We then asked to what extent the closed orbits could contribute to an identification of the experimental classes. In
the first comparison, we characterized each protagonist by a behavioral vector the entries of which were the natural symbols probabilities of the individual files (i.e., vectors of length 37). It is remarkable that on this level of simplification the similarities / dissimilarities between the classes are already expressed (see Figs. 1a–c). The characteristic stripes and peaks emerge at the same places as for the orbit analysis. This implies that the distinction among classes is to a large extent rooted in symbol probabilities. Could Drosophila courtship communication be solely based upon symbol probabilities and is the concept of closed orbits embedded in a noisy background therefore redundant?

In a second experiment, we repeated the closed orbit analysis on the surrogate data. Whereas the original closed orbits analysis corresponds to a variable-order Markov model approximation of the original data, here the underlying Markov model is of zeroth order. Of all the different models that realize a given symbol distribution, the surrogate method provides us with the most probable ones. The results (Fig. 1b) obtained for this experiment demonstrate that the basic (dis)similarities between the classes are maintained in this approximation. A detailed inspection of the three similarity matrices displayed in Fig. 1, however, confirms that the closed orbits not only may serve as the substrate for the communication among individuals, but also enhance the distinguishability between classes. To see this, note that the submatrices composed of non-fruitless vs. fruitless experiments are most clearly worked out in the original data. A network analysis of the symbol transition network corroborates these findings. Moreover, the detailed analysis of surrogate data yields that significantly less and different closed orbits are detected if compared to the original data (see Fig. 2). This implies that a large portion of the orbits found in the original data are a priori unlikely ones. Therefore, an additional, nonaccidental, structure, which can be interpreted as a grammar, underlies Drosophila courtship behavior.

Figure 2: Original files (blue) contain substantially more orbits than surrogate files (red). Estimated 30 percent surrogate orbits are not found in the original files. a) Histogram of the number of (different) closed orbits found in 120 surrogate runs, where each file was shuffled individually. b) Maximal/minimal number of orbits of a given length found across 120 surrogate runs.

General observations from the similarity matrix $M(\beta = 0)$ of the original data (see Fig. 1a) are that male Drosophila vary their behavior strongly with their partner class (no peaks of similarity within the columns indexed by 5, 7, 9), and that females vary their behaviors to a much lesser extent. For example, the behaviors of mature and mated females are rather similar, but dissimilar from the immature. This culminates in the marked distinction of the normal male behavior towards mature and immature females (columns 7 and 9), against all other behaviors. The largest contrast is obtained between columns 7 and 8, which highlights the special role of courtship between mature females and males. Dark fields adjacent to the diagonal indicate the large behavioral distance between male/female protagonists from the same experiment. Whereas female Drosophila do not discriminate particularly between fruitless and normal males, when in contact with fruitless males, normal males switch to female behavior. The high similarities displayed at the lower left corner of $M(\beta = 0)$ (indices $[1, \ldots, 4]$) express the lack of behavior adaptation by, or caused by, the fruitless mutants. The minima of similarity express the large behavioral distance between male and female Drosophila behaviors.

3. Conclusions

The performed novel analysis provides hard evidence that in this refined description, group behavior can be identified, the characterization of which is in agreement with more crude measures of behavior and with generally accepted observations by experimenters. A detailed analysis using surrogate data uncovers the presence of a nontrivial grammar in Drosophila courtship, offering a large bandwidth for pre-copulation courtship communication. Individual animals, in particular males, seem to widely use this expression space. This indicates that during courtship, more than just group membership is exchanged. Potentially transferred information range from details on the sender’s genetic configuration to the encoding of situation-dependent decisions (like the better mate vs. no mate trade-off). Our method is general enough to be applicable to a wide variety of comparative behavioral, or behavioral neurogenetics, studies that critically depend upon a sufficiently detailed quantitative analysis.

References


Abstract—This article observes a characteristic of an oscillatory associative memory that consists with a piece-wise linear hysteresis neuron. The system is one of dynamical associative memories that memorizes desired pattern as a oscillatory state. When some desired patterns are stored in the system, the system exhibits that the output wanders between some desired patterns. However, if each desired memory is orthogonal vector, the system exhibits a simple periodic output sequence.

1. Introduction

Recently, a system that can treat dynamical information, is attracted to great attention[2]-[5]. The reason why such system receives great attention, is that such dynamical information processing function can be found in biological neural networks. Especially, we think that synchronization phenomenon plays an important role for signal processing in brain.

In our previous works, we proposed Simple Hysteresis Network ( abbr. SHN )[7]. The SHN has a quite simple structure, namely, all connection has uniform value. We clarified that even such simple system exhibits periodic attractors and chaotic attractors. The SHN, however, cannot generate desired output sequence. In order to control to generate a desired output sequence, we introduce a dynamical associative memory. The each information in the dynamical associative memory is represented by phase different. Our proposed system can be regarded as a coupled oscillators system. We can observe various kinds of interesting phenomena from coupled oscillators system. In order to use such rich phenomena as information, we consider the dynamical associative memory. If the rich phenomena are corresponded to the information, we can develop a novel information processing system.

2. Hysteresis Neuron

First, we consider a hysteresis neuron[8][9]. The objective hysteresis neuron is described as

\[
\begin{align*}
\frac{dx(t)}{dt} &= -x(t) + e(y(t)), \\
y(t) &= h(x(t)) = \begin{cases} +1 & \text{for } x > -1, \\ -1 & \text{for } x < +1, \end{cases}
\end{align*}
\]  

where \( x(t) \in \mathbb{R} \) is a state variable, \( y(t) \in \{+1, -1\} \) is an output, and \( e(y(t)) \in \mathbb{R} \) corresponds to a feedback parameter. \( h(x(t)) \) is a bipolar piecewise linear hysteresis as shown in Figure 1.

Figure 1: Normalized bipolar hysteresis

The state variable of (1) varies toward the equilibrium that denotes \( e(y(t)) \). Since the equilibrium point is only depends on its output, the equilibrium point becomes a constant while the output does not change. If the equilibrium point does not exist on the hysteresis branch, the trajectory reaches the threshold of the hysteresis before it converges to the equilibrium point. When the trajectory hits the threshold of the hysteresis, the output changes its sign, and the corresponding equilibrium point changes too. This hysteresis neuron can be regarded as a relaxation oscillator whose oscillation frequency can be controlled in the feedback parameter.

Here, we consider the case where some hysteresis neurons are coupled as

\[
\tau_i \frac{dx_i(t)}{dt} = -x_i(t) + \sum_{j=1}^{N} w_{ij} y_j(t),
\]  

where \( \tau_i > 0 \) is a time constant, and \( N \) is the number of neurons. \( w_{ij} \in \mathbb{R}(i \neq j) \) is a coupling coefficient, \( w_{ii} \) is a self feedback parameter. For simplicity, we consider all time constants have the same value, hereafter.

The equilibrium point \( p_i(t) \) is given from Eqn.(2) by

\[
p_i(t) = \sum_{j=1}^{N} w_{ij} y_j(t).
\]  

If \( p_i(t) y_i(t) \leq 0 \), for some \( i \)
Figure 2: The wave form from two hysteresis neuron system. The wave form of (a) is observed in the case where the system has an excitatory connection, and (b) is observed in the case where the system has an inhibitory connection.

is satisfied, a stable output vector does not exist in the system and the output must oscillate[1].

We suppose that the connection coefficient has the following relation:

\[
\begin{align*}
|w_{ii}| & > \sum_{j=1,j\neq i}^N |w_{ij}|, \\
\sum_{j=1,j\neq i}^N |w_{ij}| & < -1, \forall i.
\end{align*}
\]

Namely, the connection coefficient matrix has diagonally dominant. If the system adopts such assumption (5), the condition (4) is satisfied. Therefore, this system must exhibit oscillatory state.

The fundamental frequency is influenced by the feedback parameter. The coupling parameters play a role as frequency modulation. Figure 2 shows typical wave forms from the two neuron system. In this figure, the upper two wave forms correspond to the state variable, \(x_1\) and \(x_2\), respectively. The lower two wave forms correspond to the output, \(y_1\) and \(y_2\), respectively.

Both results are observed into the following initial configurations. The initial value is adjusted as

\[
\begin{align*}
(x_1(0),y_1(0)) &= (+1,+1) \\
(x_2(0),y_2(0)) &= (0,+1)
\end{align*}
\]

In this case, the parameters configuration is

\[
\begin{pmatrix}
w_{11} & w_{12} \\
w_{21} & w_{22}
\end{pmatrix} = \begin{pmatrix}
-4 & 1 \\
1 & -4
\end{pmatrix}
\]

In this case, the cross-connection coefficients have a positive value. Namely, these coefficients can be regarded as excitatory connections. In this case, the system exhibits in-phase synchronization. On the other hand, the parameters configuration of Fig.2(b) is

\[
\begin{pmatrix}
w_{11} & w_{12} \\
w_{21} & w_{22}
\end{pmatrix} = \begin{pmatrix}
-4 & -1 \\
-1 & -4
\end{pmatrix}
\]

In this case, the cross-connection coefficient has a negative connection coefficient, that is an inhibitory connection. In this case, the system exhibits opposite-phase synchronization.

Figure 3: The behavior of the oscillatory associative memory which stores only one desired pattern, such that \((+1,+1,+1,+1,-1,-1,-1,-1)\).

3. Dynamical Associative Memory

In this section, we propose a dynamical associative memory whose output exhibits oscillatory state. By using a Hebbian rule, the connection coefficient is determined as

\[
W_{ij} = \frac{1}{m} \sum_{k=1}^m S_i^k S_j^k,
\]

where \(S_i^k \in [+1,-1]\) is \(i\)-th element of \(k\)-th desired pattern, and \(m\) denotes the number of desired patterns.

The network can store and retrieve oscillatory output sequence. We demonstrate the behavior of output when only one desired pattern is stored in the system which contains eight hysteresis neurons. First, we select a desired pattern as \(S \equiv (+1,+1,+1,+1,-1,-1,-1,-1)\). In this case, the system exhibits the output sequence as shown in Figure 3. Figure 3(a) shows the output sequence which generates from the system. The horizontal axis denotes time evolution, and the vertical axis denotes the output of each neuron. White pixel corresponds to the positive output, and black pixel corresponds to the negative output. The initial value of the output vector is given as \(y(0) \equiv (+1,-1,+1,-1,+1,-1,-1,+1)\). Figure 3(b) shows
The behavior of the oscillatory associative memory which stores only one desired pattern, such that \( (+1, +1, +1, +1, +1, +1, +1) \).

A direction cosine between the desired pattern and the output vector such as

\[
direction\ cosine = \frac{S \cdot y}{|S||y|}
\]  

(10)

If the direction cosine indicates +1 or −1, the system retrieves the desired pattern. In this case, the system retrieves the output sequence which consists with the desired pattern. Next, we demonstrate another example, and the behavior shows in Figure 4. The desired pattern is selected as \( S \equiv (+1, +1, +1, +1, +1, +1, +1) \). In this case, the transition time is longer than the case of Figure 4.

Finally, we examine two case where the system contains some kinds of desired patterns. First case is that the desired memories have an orthogonal relation. Figure 5 shows the output sequence from eight neurons system with three kinds of desired patterns. The three desired patterns are as follows:

- \( S^1 \equiv (+1, +1, +1, +1, -1, -1, -1) \)
- \( S^2 \equiv (-1, +1, -1, -1, +1, +1, +1) \)
- \( S^3 \equiv (-1, -1, +1, +1, -1, -1, +1) \)

The hamming distance between each desired memory is shown in Table 1. Namely, each desired memory is orthogonal vector. In this case, the system converges to a periodic output sequence which consists with two desired memories \( S^1 \) and \( S^3 \). Note that this output sequence does not contain the desired memory \( S^2 \).

Next, we show another result. The three desired patterns are as follows:

- \( S^1 \equiv (-1, +1, +1, +1, -1, -1, -1) \)
- \( S^2 \equiv (-1, +1, +1, -1, -1, +1, +1) \)
- \( S^3 \equiv (-1, -1, +1, -1, +1, +1, +1) \)

In this case, the hamming distance between each desired memory is shown in Table 3. In this case, the system exhibits that the output wanders in desired patterns \( S^1 \), \( S^2 \) and \( S^3 \) in turn. When the desired memories do not have orthogonal relation, the

Table 1: The Hamming distance between each desired memory.

<table>
<thead>
<tr>
<th></th>
<th>S^1</th>
<th>S^2</th>
<th>S^3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S^1</td>
<td>-</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>S^2</td>
<td>4</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td>S^3</td>
<td>4</td>
<td>4</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 2: The retrieving rate of the desired memories in the output sequence.
In order to confirm this phenomenon, we carry out such numerical simulation 100 times. The desired memories of each simulation are three, and all desired memories do not have an orthogonal relation. Table 4 shows each rate. As a result, almost all output sequences contain all desired memories with wondering motion.

4. Conclusions

In this article, we observed the synchronization phenomena from the oscillatory associative memory. The proposed dynamical associative memory consists of the hysteresis neuron which can be regarded as a relaxation oscillator. First, we clarified that the system exhibits in-phase synchronization oscillation when the cross-connection coefficient denotes excitatory connection. Also, the system exhibits opposite-phase synchronization when the cross-connection coefficient denotes inhibitory connection. Based on this result, we proposed the oscillatory associative memory. The oscillatory associative memory exhibits an interesting wondering motion which is depended on the kind of desired memories. The theoretical analysis is one of our future problems.

References


<table>
<thead>
<tr>
<th></th>
<th>$S'_1$</th>
<th>$S'_2$</th>
<th>$S'_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S'_1$</td>
<td>-</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$S'_2$</td>
<td>2</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td>$S'_3$</td>
<td>3</td>
<td>3</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 3: The hamming distance between each desired memory.

| All desired memories are contained | 93% |
| 2 desired memories are contained | 0%  |
| Only 1 desired memory is contained | 7%  |
| No desired memory is contained   | 0%  |

Table 4: The retrieving rate of the desired memories in the output sequence.
Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition

Yuko OSANA†

†School of Computer Science, Tokyo University of Technology
1404-1 Katakura-cho, Hachioji, Tokyo, 192-0982, Japan
Email: osana@cc.teu.ac.jp

Abstract—In this paper, we propose an Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition (IHCAMSL-MW). The proposed model is based on a Hetero Chaotic Associative Memory for Successive Learning with give up function (HCAMSL) and a Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition (HCAMSL-MW). In the proposed IHCAMSL-MW, the learning process and the recall process are not divided. When an unstored pattern is given to the network, the IHCAMSL-MW can learn the pattern successively and its storage capacity is larger than that of conventional HCAMSL/HCAMSL-MW.

1. Introduction

Recently, neural networks are drawing much attention as a method to realize flexible information processing, and many models have been proposed[1]–[4]. In these models, the learning process and the recall process are divided, and therefore they need all information to learn in advance. However, in the real world, it is very difficult to get all information to learn in advance. So we need the model whose learning and recall processes are not divided. As such model, Grossberg and Carpenter proposed ART (Adaptive Resonance Theory)[5]. However, the ART is based on the local representation, and therefore it is not robust for damage. While in the field of associative memories, some models have been proposed[6]–[8]. Since these models are based on the distributed representation, they have the robustness for damaged neurons. However, their storage capacity is very small because their learning processes are based on Hebbian learning.

Recently, we have proposed a Hetero Chaotic Associative Memory for Successive Learning with give up function (HCAMSL)[9][10] and a Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition (HCAMSL-MW)[11] in order to improve the storage capacity.

In this paper, we propose an Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition (IHCAMSL-MW) by integrating the conventional HCAMSL and HCAMSL-MW. In the proposed IHCAMSL-MW, the learning process and the recall process are not divided. When an unstored pattern is given to the network, the IHCAMSL-MW can learn the pattern successively. Moreover, the storage capacity of the proposed IHCAMSL-MW is larger than that of the conventional HCAMSL/HCAMSL-MW.

2. Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition

2.1. Outline of IHCAMSL-MW

Here, we explain the outline of the proposed Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition(IHCAMSL-MW). The proposed IHCAMSL-MW has three stages: (1) Pattern Search Stage, (2) Distributed Pattern Generation Stage and (3) Learning Stage.

When an unstored pattern set is given to the network, the proposed IHCAMSL-MW distinguishes an unstored pattern set from stored patterns and can learn the pattern set successively. When a stored pattern set is given, the IHCAMSL-MW recalls the patterns. When an unstored pattern set is given to the network, the IHCAMSL-MW changes the internal pattern for the input pattern set by chaos and presents other pattern candidates (we call this the Pattern Search Stage). When the IHCAMSL-MW can not recall the desired patterns, the distributed pattern is generated by the multi-winners competition[12] (Distributed Pattern Generation Stage), and it learns the input pattern set as an unstored pattern set (Learning Stage).

2.2. Structure of IHCAMSL-MW

The proposed IHCAMSL-MW is a kind of the hetero-associative memories. Figure 1 shows a structure of the IHCAMSL-MW. This model has two layers; an Input-Output Layer (I/O Layer) composed of conventional neurons and a Distributed Representation Layer (DR Layer) composed of chaotic neurons[13]. In this model, there are the connection weights between neurons in the Distributed Representation Layer and the connection weights between the Input-Output Layer and the Distributed Representation Layer. As shown in Fig.1, the Input-Output Layer has plural parts. The number of parts is decided depending on the
In this model, when a pattern set is given to the Input-Output Layer, the internal pattern corresponding to the input patterns is formed in the Distributed Representation Layer. Then, in the Input-Output Layer, an output pattern set is generated from the internal pattern. The IHCAMSL-MW distinguishes an unstored pattern set from stored patterns by comparing the input patterns with the output patterns.

In this model, the output of the neuron $i$ in the Distributed Representation Layer at time $t+1$, $x_i^D(t+1)$ is given by the following equations.

$$x_i^D(t+1) = \begin{cases} \phi_D(x_i^D(t+1)), & \text{when a new pattern set is given} \ (1) \\ \phi_D(x_i^D(t+1) + \eta_i(t+1) + \xi_i(t+1)), & \text{otherwise} \end{cases}$$

In eqs.(1)-(4), $M$ is the number of neurons in the Input-Output Layer, $v_{ij}$ is the connection weight between the neuron $j$ in the Input-Output Layer and the neuron $i$ in the Distributed Representation Layer, $N$ is the number of neurons in the Distributed Representation Layer, $A_j(t)$ is the external input to the neuron $j$ in the Input-Output Layer at the time $t$, $w_{ij}$ is the connection weight between the neurons $i$ and $j$ in the Distributed Representation Layer, $\alpha(t)$ is the scaling factor of the refractoriness at the time $t$, $k_1$, $\eta_n$, $k_2$ are the damping factors. $\phi_D(\cdot)$ is the following output function:

$$\phi_D(u_i) = \tanh(u_i/\varepsilon)$$

where $\varepsilon$ is the steepness parameter.

The output of the neuron $j$ in the Input-Output Layer at the time $t$, $x_j^{IO}(t)$ is given as follows.

$$x_j^{IO}(t) = \phi^{IO}\left(\sum_{i=1}^{N} v_{ij} x_i^D(t)\right)$$

$$\phi^{IO} = \begin{cases} 1, & u \geq 0 \\ -1, & u < 0 \end{cases}$$

### 2.3. Pattern Search Stage

In the Pattern Search Stage, when an input pattern set is given, the IHCAMSL-MW distinguishes the pattern set from stored patterns. When an unstored pattern set is given, the IHCAMSL-MW changes the internal pattern for the input pattern by chaos and presents the other pattern candidates. Until the IHCAMSL-MW recalls the desired patterns, the following procedures are repeated. If the IHCAMSL-MW can not recall the desired patterns, when the stage is repeated certain times, the IHCAMSL-MW finishes the stage.

(1) **Pattern Assumption**

In the proposed IHCAMSL-MW, only when the input patterns are given to all parts of the Input-Output Layer, the patterns are judged. When the input pattern $A(t)$ is similar to the recalled pattern $x^{IO}(t)$, the IHCAMSL-MW can assume that the input pattern is one of the stored patterns. The IHCAMSL-MW outputs the pattern formed by the internal pattern in the Distributed Representation Layer. The similarity $s(t)$ is defined by:

$$s(t) = \frac{1}{M} \sum_{j=1}^{M} g(A_j(t), x_j^{IO}(t))$$

$$g(a, b) = \begin{cases} 1, & a = b \\ 0, & a \neq b \end{cases}$$

The IHCAMSL-MW regards the input patterns as a stored pattern set, when the similarity rate $s(t)$ is larger than the threshold $s^{th}$ ($s(t) \geq s^{th}$).

(2) **Pattern Search**

When the IHCAMSL-MW assumes that the input patterns is an unstored pattern set, the IHCAMSL-MW changes the internal pattern $x^{D}(t)$ for the input patterns by chaos and presents the other pattern candidates.

In the chaotic neural network, it is known that dynamic (chaotic) association can be realized if the scaling factor of the refractoriness $\alpha(t)$ is suitable[13][14]. Therefore, in the proposed model, $\alpha(t)$ is changed as follows:

$$\alpha(t) = (\alpha_{\max}(t) - \alpha_{\min}(t))(1 - s(t)) + \alpha_{\min}(t)/\alpha_{DIV}$$

$$\alpha_{\max}(t) = M v_{\max} + N w_{\max}$$

$$v_{\max} = \max[|v_{11}|, \cdots, |v_{ip}|, \cdots, |v_{NN}|]$$

$$w_{\max} = \max[|w_{11}|, \cdots, |w_{ip}|, \cdots, |w_{NN}|]$$

where $\alpha_{\min}$ is the minimum of $\alpha$, $\alpha_{\max}$ is the maximum of $\alpha$ at the time $t$, $s(T)$ is the similarity between the input
pattern and the output pattern at the time \( T \) (the time when the Pattern Search Stage started), \( \alpha_{DIV} \) is constant.

### 2.4. Distributed Pattern Generation Stage

In the Distributed Pattern Generation Stage, the distributed pattern corresponding to the input pattern is generated by the multi-winners competition[12].

#### Step 1 : Calculation of Outputs of Neurons in I/O Layer

When the input pattern \( A_i(t) \) is given to the Input-Output Layer, the output of the neuron \( j \) in the Input-Output Layer \( x_j^{IO} \) is given

\[
x_j^{IO} = S_f(A_j(t))
\]

where \( S_f(\cdot) \) is the ramp function and is given by

\[
S_f(u) = \begin{cases} u, & u > 0 \\ 0, & u \leq 0. \end{cases}
\]

#### Step 2 : Calculation of Initial Outputs of Neurons in DR Layer

The output of the neuron \( i \) in the Distributed Representation Layer \( x_i^{D(0)} \) is calculated by

\[
x_i^{D(0)} = \phi^D \left( \sum_{j=1}^{M} v_{ij} x_j^{IO} - \theta_i \right)
\]

where \( v_{ij} \) is the connection weight from the neuron \( j \) in the Input-Output Layer to the neuron \( i \) in the Distributed Representation Layer, \( \theta_i \) is the threshold of the neuron \( i \) in the Distributed Representation Layer and \( M \) is the number of neurons in the Input-Output Layer. where \( T \) is the steepness parameter in the sigmoidal function.

#### Step 3 : Competition between Neurons in DR Layer

The competition dynamics is given by the following equation.

\[
x_i^{D} = C_f(u_i - \theta_i)
\]

\[
\tau \frac{d}{dt} u_i = -u_i + \sum_{j=1}^{N} w_{ij} x_j^{D}
\]

where \( x_i^{D} \) is the output of the neuron \( i \) in the Distributed Representation Layer, \( u_i \) is the internal state of the neuron \( i \) in the Distributed Representation Layer and \( N \) is the number of neurons in the Distributed Representation Layer.

### 2.5. Learning Stage

In the Pattern Search Stage, if the IHCAMSL-MW cannot recall the desired pattern set, it learns the input pattern set as an unstored pattern set. The Learning Stage has two phases: (1) Hebbian Learning Phase and (2) anti-Hebbian Learning Phase. In the Hebbian Learning Phase, if the signs of the outputs of two neurons are the same, the connection weight between these two neurons is strengthened. By this learning, the connection weights are changed to learn the input patterns, however the Hebbian learning can only learn a new input pattern set. In the proposed IHCAMSL-MW, the anti-Hebbian Learning Phase is employed as similar as the original CAMSL[8] and the HCAMSL[9]. In the anti-Hebbian Learning Phase, the connection weights are changed in the opposite direction in the case of the Hebbian Learning Phase. The IHCAMSL-MW can learn a new pattern set without destroying the stored patterns by the anti-Hebbian Learning.

#### (1) Hebbian Learning Phase

In the Hebbian Learning Phase, until the similarity rate \( s(t) \) becomes 1.0, the following procedures ((a) and (b)) are repeated.

(a) Update of Connection Weights

The connection weight between the Input-Output Layer and the Distributed Representation Layer \( v_{ij} \) and the connection weight in the Distributed Representation Layer \( w_{ij'} \) are updated as follows :

\[
v_{ij}^{(new)} = v_{ij}^{(old)} + \gamma_v x_i^D(t) A_j(t)
\]

\[
w_{ij'}^{(new)} = w_{ij'}^{(old)} + \gamma_v x_i^D(t) A_{ij'}(t)
\]

where \( \gamma_v \) is the learning rate of the connection weight \( v_{ij} \) in the Hebbian Learning Phase, and \( \gamma_v \) is the learning rate of the connection weight \( w_{ij'} \) in this phase.

(b) Normalization of Connection Weights

When the number of the learnings in this phase becomes more than the threshold \( T_n \) and the similarity rate \( s(t) \) is not equal to 1.0, the connection weights are normalized.

\[
v_{ij}^{(new)} = \frac{v_{ij}^{(old)}}{v_{max}}
\]

\[
w_{ij'}^{(new)} = \frac{w_{ij'}^{(old)}}{w_{max}}
\]

where \( v_{max} \) is the maximum absolute value of the connection weights \( v_{ij} \), and \( w_{max} \) is the maximum absolute value of the connection weight \( w_{ij'} \).

(c) Give Up Function

When the similarity rate \( s(t) \) does not become 1.0 even if connection weights are normalized \( T_n \) times, the IHCAMSL-MW gives up to memorize the patterns. If the IHCAMSL-MW gives up to learn the patterns, the anti-Hebbian Learning Phase is not performed.

#### (2) Anti-Hebbian Learning Phase

The anti-Hebbian Learning Phase is performed after the Hebbian Learning Phase. In this phase, the connection weights \( v_{ij} \) and \( w_{ij'} \) are changed in the opposite direction in the case of the Hebbian Learning Phase. The anti-Hebbian Learning makes the relation between the patterns are learned without destroying the stored patterns.
4. Conclusions

In this paper, we have proposed the Improved Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition (IHCAMSL-MW). The proposed model is based on a Hetero Chaotic Associative Memory for Successive Learning with give up function (HCAMSL) and a Hetero Chaotic Associative Memory for Successive Learning with Multi-Winners Competition (HCAMSL-MW). In the proposed HCAMSL-MW, the learning process and the recall process are not divided. When an unstored pattern is given to the network, the IHCAMSL-MW can learn the pattern successively. We carried out a series of experiment and confirmed that the proposed IHCAMSL-MW can memorize pattern sets successively and its storage capacity is larger than that of conventional HCAMSL/HCAMSL-MW.

References

The parameters of the Ornstein-Uhlenbeck neuronal model

Petr Lansky†, Pavel Sanda†, Jufang He‡

†Institute of Physiology, Academy of Sciences of the Czech Republic, Prague, Czech Republic
‡Department of Rehabilitation Sciences, The Hong Kong Polytechnic University, Hong Kong
Email: lansky@biomed.cas.cz

Abstract—Five parameters of one of the most common neuronal models, the Ornstein-Uhlenbeck process, were estimated on the basis of intracellular recording. These parameters are classified into two categories. The membrane time constant, the resting potential and the firing threshold characterize the neuron itself. The remaining two parameters characterize the neuronal input. The intracellular data used for the estimation were collected during spontaneous firing. Two methods for the estimation were applied, the regression method and the maximum-likelihood method. Both methods permit to estimate the input parameters and the membrane time constant in a single interspike interval. We found that, at least in our example, the regression method gave more consistent results than the maximum-likelihood method. The model neuron, as deduced from the determined parameters, works in a subthreshold regime. The spikes are generated when the membrane depolarization is not an inherent part of the model as in more complex leaky integrate-and-fire models. Thus the model is fully described by equation (1) and the process of input stochasticity into the deterministic leaky integrate-and-fire model is simply by assuming an additional noise term. If the noise is not further specified, but assumed to be Gaussian and white, then the model is known as an Ornstein-Uhlenbeck model. For details on this neuronal model see [1].

We aimed to study the estimation methods in the Ornstein-Uhlenbeck model, their stability and reproducibility. We restricted the study on a single neuron under spontaneous activity conditions. Extended version of this work will appear in [2].

2. Model and its properties

The Ornstein-Uhlenbeck model of membrane depolarization is given by the stochastic differential equation, which can be rewritten in a form often seen in engineering applications

\[ \frac{dX(t)}{dt} = -\beta(X(t) - x_0) + \mu + \sigma \xi(t), \quad X(0) = x_0, \quad (1) \]

where the white noise \( \xi(t) \) is a formal derivative of the Wiener process with respect to time, \( \beta > 0 \) characterizes the spontaneous decay of the membrane depolarization in the absence of input to the resting level \( x_0 \). The drift coefficient \( \mu \) reflects the local average rate of displacement due to the neuronal input and local variability is represented by the infinitesimal variance \( \sigma \) (the variability of the neuronal input). The spikes are generated when the membrane depolarization \( X(t) \) reaches for the first time the firing threshold \( S \), which is an additional parameter. Then, the depolarization is reset to the resting level, \( x_0 \) and the process of input “integration” starts anew. We should keep in mind that also the reset level, \( x_0 \), represents an additional parameter of the model. Thus the model is fully described by equation (1) with its five parameters: \( \beta, \mu, \sigma, S \) and \( x_0 \).

Due to the complete reset in defining ISI and due to the constant input \( \mu \), the ISIs form a renewal process, which means that ISIs are independent and identically distributed random variables.

For a fixed time \( t \), \( X(t) \) given by (1) is a Gaussian random variable. In absence of the threshold \( S \) and if \( \sigma \) tends to zero, we can solve the differential equation (1). The solution is identical with the mean value of the stochastic de-
The position of the asymptotic depolarization 
\[ E(X(\infty)) = x_0 + \frac{\mu}{\beta} \]

The description of the process via equation (1) is appar-
ently an intuitive extension of the deterministic approach.
Its advantage is in giving a method for a computer sim-
ulation of the process sample trajectories. The simplest
discrete-time approximation of (1) is a stochastic analogue
of the Euler scheme for ordinary differential equations,
which was the procedure applied to simulate the membrane
depolarization in this study. Apparently, the parameters
\[ S \]
\[ \sigma \]
\[ \mu \]
\[ x_0 \]
are input parameters and thus are assumed to
The membrane depolarization is recorded between the
generation of spikes. To specify the firing threshold and re-
set level seems to be a simpler task than to estimate the remaining parameters of the model. We should simply record what was the reset after the end of an action potential and what was the final value of the depolarization when it started. However, we will see that the situation is not so simple and also these two parameters need to be estimated. It is not a typical task when treating the diffusion processes and it comes out from the fact that our data are repeatedly reset when reaching the firing threshold.

The aim of this article is primarily determination of the values of the parameters \( \beta, \sigma \) and \( \mu \). This is a standard task investigated in the theory of statistical inference for stochastic diffusion processes [3]. A method of estimating these parameters was proposed more than two decades ago [4]. The question is whether these parameters are stable over a long period or whether they vary in short time ranges. Whereas \( \sigma \) and \( \mu \) are input parameters and thus are assumed to change whenever the input to a neuron has changed, model (1) assumes that \( \beta \) is a property of the membrane (in the same way as \( S \) and the reset level) and these three intrinsic parameters should be stable. However, these are only assumptions which have never been confirmed. Thus, initially we estimate the parameters separately for each ISI.

3. Methods

3.1. Estimations from interspike interval(s)

The records of the depolarization within single ISI permit us to estimate \( \beta, \mu \) and \( \sigma \) by the maximum likelihood method. Formula (2) suggests that the method of moments can also be used. Then, we minimize the functional with respect to the parameters \( \beta \) and \( \mu \) by a regression method.

As we estimated the parameters for each ISI separately, we also got some information about the dependency of the values on the lengths of ISI and their position in the record. The global record can be characterized by representative median values of \( \beta_M, \mu_M \) and \( \sigma_M \). Also the median threshold \( S \) and the median reset value \( x_0 \) were estimated.

3.2. Intracellular recordings and data collection

Guinea pigs served as subjects for the intracellular recording experiments. Throughout the recording, an electrocorticograph was monitored to assess the level of anaesthesis. A midline incision was made in the scalp and a craniotomy was performed to enable vertical access to the MGB in the right hemisphere. The experimental procedures were approved by the Animal Subjects Ethics Sub-Committee of The Hong Kong Polytechnic University.

Upon penetrating the membrane of a cell, the electrode detected the negative membrane potential. After amplification, the membrane potential as well as the auditory stimulus were stored in the computer with the aid of commercial software (AxoScope, Axon). Single neuron data were selected for this article. The membrane potential was recorded (in 100 mV) with time step \( h = 0.00015 [s] = 0.15 [ms] \), for period \( 0 - 501 [s] \). For the purpose of this study we selected only ISIs which were entirely outside the stimulation period.

3.3. Spikes detection, determination of \( S \) and \( x_0 \)

From visual inspection of the data it is clearly difficult to decide where exactly to start and to end the spikes. It follows from this inspection that for detected spikes the width of the spike as well as the voltage where ISI starts, are not always the same. This is in contrast with the assumptions of model (1). Determining \( x_0 \) by the minimum voltage after detected spike failed due to the large fluctuations of these values. The final solution, which was adopted, was that all data were transformed by a moving average (over six data points) and the minimum in “the valley” after a spike is considered to be start of an ISI. The value of the voltage at the minimum after the spike is taken as an estimate of \( x_0 \).

Defining the end of each ISI was not so problematic and we took the point 0.01005 [s] before detected spike. For threshold determination, we took the last point with decreasing depolarization before the spike detection. This value is considered as an estimate of the parameter \( S \).
4. Results

4.1. Input parameter and the membrane time constant

Using the above described procedure we identified all ISIs and estimated the parameters for each of them. The ISIs appear stable in time. The shape of histogram of ISIs suggests that the ISIs are generated in accordance with the exponential distribution. From the simulations made with estimated parameters, it appears that the estimated asymptotic voltages, $\mu/\beta$, coincide well with the data. On the other hand, in the first part of the trajectories, the real data is a bit faster in reaching the asymptote than simulated trajectory. The visual impression was confirmed by the method which aimed to check the fit of the model to the data. The main difference between two estimation methods is in the period just after the spike generation. This could be due to a violation of the model assumptions (for example, hyperpolarization) and possibly the method of moments could be more robust against this violation. In both cases there is a systematic hump after the origin, but for the regression method it is much smaller, so the regression method works better.

An important question is dependency of the estimated parameters on the length of ISIs. The only dependency we can expect that, if the input to neurons changes with the experiment, then $\mu$ could get smaller for longer ISI. Otherwise, $\mu$ and $\beta$ should keep stable and independent on the length of ISI. The results obtained by the regression method are independent of ISI which is not the case of the estimates obtained by the maximum likelihood method.

Similarly, we investigated the estimated value of $\sigma$ and found it independent of the ISI length. As the regression method was superior to the maximum likelihood method for estimation of $\mu$ and $\beta$, the values from the regression are considered further on.

Finally we calculated the central characteristics of the estimated parameters. Median value of the noise amplitude was $\bar{\sigma}_M = 0.013505\ [V/\sqrt{s}]$. Median values of $\mu$ and $\beta$ were by regression method, $\hat{\mu}_M = 0.2846\ [V/s], \hat{\beta}_M = 25, 8042\ [1/s]$, for the maximum likelihood method the obtained values were $\bar{\mu}_M = 0.4606\ [V/s], \bar{\beta}_M = 43, 5068\ [1/s]$. Due to the symmetry of histograms, the averages and the medians were practically the same. We should notice that both methods give almost identical asymptotic depolarization $\mu/\beta$, for regression 0.0110 [V] and for maximum likelihood 0.0106 [V]. We can calculate the membrane time constant, which from the regression method yields after inverting the estimate of $\beta$, the value of 38.8 ms.

4.2. Reset, threshold, asymptotic depolarization

For each ISI we estimated the initial value of the depolarization after a spike and the firing threshold (the last value before the spike is generated).

The initial values were more variable than the thresholds $S$. The reason may be that the spike was not in principle generated at the time when the voltage was at its highest level during the ISI and this will be discussed later. The median value of the initial depolarization is $x_0 = -73.92\ [mV]$. The median threshold value is $S = -61.0\ [mV]$. It implies that, in average, the firing threshold is about $13\ [mV]$ above the initial depolarization. It is larger than approximately $11\ [mV]$ which is the level of asymptotic depolarization as it comes out from the parameters estimation. The normality of both determined parameters, $x_0$ and $S$, was rejected.

Finally, to show that the neuron, as the estimation of the parameters suggests, is in the subthreshold (noise-driven) regimen, we compared directly the theoretical asymptotic depolarization and the corresponding firing threshold. It appeared that the difference between the threshold and the asymptotic depolarization is almost always positive. If we investigate some kind of two-standard-deviations envelope around the asymptotic depolarization, then we get below zero.

5. Discussion

5.1. Spontaneous activity

We selected for statistical treatment only the unstimulated activity of the neuron. The detected ISIs are exponentially distributed which suggests that they are generated in accordance with a Poisson process. The spontaneous activity can be interpreted as the summation of two processes: (i) an intrinsic process which implies firing due to the noise, and (ii) an extrinsic process, which induces firing due to the uncontrolled occurrence of effects either from the environment or other neurons. Of course, Poisson process, as any other model, is an approximation which can be always questioned, in this case for example by the existence of the absolute refractory period, but such objections are marginal at this level of description.

From the point of view of the Ornstein-Uhlenbeck neuronal model, the Poisson process corresponds to the situation in which the signal is so weak that the asymptotic depolarization is far below the firing threshold, which has to be true with respect to the amplitude of noise. In other words, for model (1) this type of activity is predicted if the firing threshold $S$ is far above the asymptotic depolarization, $\mu/\beta$, given by equation (2) with respect the asymptotic variance which is controlled by $\sigma$.

5.2. Model assumptions

There are many assumptions of the model which are disputable and can be true only in idealization. First of all, it is the constancy of the input parameters $\mu$ and $\sigma$ over each ISI. We selected the spontaneous activity as an experimental material, for which such a constancy can be expected. Another assumption is the Gaussian white noise on the neuronal input. Next assumption, violated in our observations, was the first-passage time inducing the generation
of a spike. Probably, we registered the membrane depolarization in other location than the spikes are generated. The traditional assumption of the model is the fixed value of the reset depolarization. This assumption represents an oversimplification. We found that this assumption was also violated, but for the model performance the effect is negligible. Despite that any assumption of the model can be made questionable, as it is only a model, we may conclude that the data are consistent with the model.

5.3. Intrinsic parameters

In model (1) are three intrinsic parameters, \( \beta \), \( S \) and \( x_0 \), and for different neurons have been reported different values of them, even without modeling concept. However, the intrinsic parameters were not in the center of our interest despite they play their important role for the model performance. In general, the values of intrinsic parameters obtained in this article are consistent with the values found in literature.

5.4. Input parameters

Completely different situation comes with the input parameters. The attempts to estimate them were up to now rare and based on additional assumptions. These attempts [5][6] were based on ISIs statistics. It means that for estimation of the parameters a sample of several hundreds of ISIs is necessary. Our method permits to estimate the input parameters in a short time window (in a single ISI). It appeared that the regression method was superior to the maximum likelihood. There might be several reasons for this effect. The first one can be the above mentioned violation of the model assumptions and thus that the maximum-likelihood may be more sensitive to these discrepancies between the model and the data. The second reason is that the maximum likelihood estimates are discretized version of continuous sampling theory. The asymptotic depolarization \( \mu/\beta \) was estimated very well by both methods, better than the parameters \( \mu \) and \( \beta \) separately. The reason is that the membrane potential was almost permanently at the asymptotic level and in this situation the estimation of individual parameters is less precise.

5.5. Noise

The results suggest that our neuron was firing in the noise activated regimen, in other words, that in the absence of the noise it would remain silent. This corresponds very well to the fact that the driving signal is small and the neuron fires only due to the stochastic fluctuation of the membrane depolarization. Theoretical prediction of the Poissonian firing in the subthreshold regime of the Ornstein-Uhlenbeck neuronal model is well known for a long time and here the prediction and data estimation fits perfectly together.

Acknowledgements

The authors thank to P.W.F. Poon for initiating and permanent support for this work and to P.E. Greenwood for stimulating discussions. This work was supported by Academy of Sciences of the Czech Republic Grant (Information Society, 1ET400110401), AV0Z50110509, Center for Neurosciences LC554.

References

Oscillation Mechanism in Cyclic Coupled Neurons

Hiroyuki Kitajima†, Takashi Ishii and Tetsuo Hattori
Faculty of Engineering, Kagawa University
2217-20 Hayashi Takamatsu, Kagawa 761-0396, Japan
Email: ?kitaji@eng.kagawa-u.ac.jp

Abstract—We consider unidirectionally coupled ring networks of neurons with inhibitory connections. It is known that when a number of inhibitory neurons is even the system never has stable oscillatory modes. However, we observed oscillatory modes in such a case. In this paper we investigate the relationship between these oscillatory modes and the unstable periodic solutions generated by the Hopf bifurcations.

1. Introduction

Systems of coupled oscillators have been widely studied as models of information processing in the brain [1], animal locomotion [2], generating nonlinear phenomena such as chaotic itinerancy [3] and on-off intermittency [4, 5], and so on. In particular, we are interested in neuron models, because many investigators confirm that oscillatory dynamics of neural activity and its synchronization play an important role in information processing in the brain.

In this paper, we consider unidirectionally coupled ring networks of neurons with inhibitory connections. It is said that the essential dynamics of some biological central pattern generators can be captured by a model consisting of $N$ neurons connected in a ring structure [6–8]. In such a system, it is known that when a number of inhibitory neurons is odd and even, the system has stable and unstable oscillatory modes, respectively [9,10]. However, we found out the existence of oscillatory modes as a quasi-attractor in a system of even inhibitory neurons [11]. In this study we investigate the relationship between these oscillatory modes and the unstable periodic solutions generated by the Hopf bifurcations. We also suggest a method of calculating unstable periodic solutions in cyclic coupled cells.

2. System Equation

The system equation is described as

$$\frac{dx_i}{dt} = -x_i - cf(x_{i+1}) \quad (i = 1, 2, \cdots, n, \ x_{n+1} \equiv x_1)$$

(1)

where $x_i$ is the state of neuron $i$, $n$ is the number of neurons, $\tau$ is the time constant fixed as 1.0, $c$ is the coupling coefficient and $f(x)$ is the output function given by $\tan^{-1}(x)$. This neuron model is commonly used for controlling locomotion [12–14] and describing oscillatory phenomena [10,15].

We consider $c > 0$ in Eq. (1), thus Eq. (1) consists of only inhibitory neurons. The Jacobi matrix of Eq. (1) evaluated at the origin (one of equilibrium points) can be written as:

$$C = -I_n - cP_n$$

(2)

where $I_n$ is the identity matrix and $P_n$ is a cyclic matrix given by

$$P_n = \begin{bmatrix}
0 & 1 & \cdots & 0 \\
0 & 0 & 1 & \ddots \\
0 & & & \ddots & 0 \\
\vdots & & \ddots & \ddots & 0 \\
1 & \cdots & 0 & 0 & 0
\end{bmatrix}$$

(3)

The eigenvalues of $C$ are analytically obtained as

$$g(1) = -1 - c$$
$$g(a) = -1 - ca$$

(4)

where $a = e^{i2\pi/n}$. Then, we obtain distribution of the eigenvalues on the complex plane shown as Fig. 1 ($c = 1.5$ and $n = 10$). When we change the value of the parameter $c$ (that corresponds to the change of the diameter of the circle in Fig. 1) and a pair of the eigenvalues becomes pure imaginary, the Hopf bifurcation occurs. Such a value of imaginary numbers indicates an angular frequency of an oscillation generated by the Hopf bifurcation. Here, we define a number of occurrence of the Hopf bifurcation as $k$. In Fig. 1 $k = 1$. Thus, an angular frequency is proportionally increased to the value of $k$.

3. Method of obtaining unstable periodic solutions

The phase difference $\psi$ of oscillations generated by the $k$-th Hopf bifurcation for nearest neurons is given by

$$\psi = \frac{2\pi}{n} \times k - \pi$$

(5)
If there exists a natural number $M$ such that $M\phi = -l\pi$ where $l$ is also a natural number, then we can obtain the unstable periodic solutions by putting initial states as $x_i = x_{i+l}$ and $x_i = -x_{i+M}$ for even and odd $l$, respectively.

4. Results

4.1. Unstable Periodic Solution (UPS)

We use Fehlberg’s formula for numerical integration of Eq. (1). As an illustrated example of Sec. 3, we show the case of $n = 80$. In this case we can obtain UPSs when $k = 1, 2, 4, 5, 8$ and $10$, but here we only show $k = 1, 2, 8$ for simplicity.

The first Hopf bifurcation ($k = 1$) occurs at $c \approx 1.0031$. In this case we derive $M = 40$ and odd $l$. Thus by putting initial states as $x_i = -x_{i+40}$ ($i = 1, \ldots , 40$) we obtain the 80-phase UPS shown in Fig. 2(a). From this figure we can see that $x_1$ and $x_{41}$ are in anti-phase. This solution is one-dimensionally unstable, thus we cannot observe it even if one of the initial conditions $x_i = -x_{i+40}$ is not satisfied.

We investigate the stability of this oscillation. Figure 3 shows the value of $\lambda - 1$, where $\lambda$ is the eigenvalue ($\lambda > 1$; only one eigenvalue is over one in this case) of the Jacobi matrix of the Poincaré map evaluated at this periodic point, as a function of the coupling coefficient $c$. This solution is one-dimensionally unstable at the bifurcation point. By increasing the value of $c$ the instability is weaker and weaker, but it does not become stable until $c = 4$.

The second Hopf bifurcation occurs at $c \approx 1.0125$. The waveform of the UPS generated by this Hopf bifurcation is shown in Fig. 2(b). $M = 20$ and $l$ is odd, thus $x_1$ and $x_{21}$ are anti-phase. Initial conditions are $x_i = -x_{i+20}$ ($i = 1, \ldots , 20, 41, \ldots , 60$) and $x_i = x_{i+40}$ ($i = 1, \ldots , 40$). This solution is three-dimensionally unstable.

The eighth Hopf bifurcation occurs at $c \approx 1.315$. Figure 2(c) shows the waveform of the UPS generated by this Hopf bifurcation. We calculate that $M = 5$ and $l$ is even, thus $x_i$ and $x_{i+5}$ are in-phase. Initial conditions are $x_i = x_{i+5}$ ($i = 1, \ldots , 75$). This UPS is 5-phase.

As mentioned in Sec. 3, $k$ is proportional to the angular frequency of oscillations, so the frequency is increased in order of Fig. 2(a) to (c).

4.2. Phase Difference

We assume that a firing of the neuron described by Eq. (1) occurs when the state variables $x_i$ crosses zero as a threshold value, changing its sign from negative to positive. Then we define the phase difference $\theta_i$ between neurons $x_i$ and $x_j$: $\theta_i = \frac{t_j - \phi_i}{t_l} - t_l$ (6)

where $s$ is a number of firing for neuron $x_i$, $t_l$ and $t_{l+1}$ are respectively $s$-th and $(s + 1)$-th firing times for neuron $x_i$, and $t_l$ is firing time for neuron $x_i$ between $t^i$ and $t^{i+1}$. Using this phase difference we can investigate the relation between oscillatory modes which we found in [11] and the UPSs generated by the Hopf bifurcations.

Figures 4 and 5 show phase difference $\theta_i$ for the oscillatory mode from random initial states and UPSs, respectively. Note that $\theta_i \neq \theta_j$, thus these figures are not completely symmetrical for $i = j$. In Fig. 4(a) and (b), we observe several green dots corresponding to random initial states. At $s = 17$ (Fig. 4(c)) the stationary pattern is beginning to form. On the other hand, in Fig. 5 the patterns are already formed in the small value of $s$, because initial states have such patterns. As mentioned in Sec. 4.1, neurons $x_i$ and $x_{i+40}$, and $x_i$ and $x_{i+20}$ are anti-phase shown by red dots in Fig. 5(a) and Fig. 5(b), respectively.

Next, we define the Poincaré section as

$$\Sigma = \{ x \in R^n | x_i = 0, x_i > 0 (i = 2, \ldots , n) \}$$

Here we can calculate the Euclidean distance $D$ between the oscillatory mode and the UPS for $k = 1$ shown in Fig. 4 and 5, respectively, on the Poincaré section (Fig. 6). Form this figure we can see that

- the oscillatory solution from random initial states converges at $t \approx 650$, that corresponds to $s = 17$ in Fig. 4(c),
- the oscillatory pattern in Fig. 4(d) exists near the UPS for $k = 1$, but these are different, because $D \neq 0$.

5. Conclusion

We investigate the oscillatory stable solution and UPSs in cyclic coupled neurons. First, we show the method of calculating UPSs by setting initial states as in-phase or anti-phase. Three UPSs are shown to demonstrate effectiveness of our method. Second, we study the relationship between...
Figure 2: Waveforms of UPSs after transient time.

(a) $c = 1.004$ and $k = 1$.

(b) $c = 1.02$ and $k = 2$.

(c) $c = 1.4$ and $k = 8$.

Figure 3: Eigenvalue over one for UPS of $k = 1$ as a function of the coupling coefficient $c$.

Figure 4: Phase differences for an oscillatory mode from random initial states. Colors are used for intervals of $\theta_{ij}$ shown above. If we cannot calculate $t_j$ that corresponds to non-firing state for $j$-th neuron, or $\theta_{ij} > 1$, we use green.
the oscillatory solution and UPSs by phase difference patterns for all neurons and Euclidean distance on the Poincaré section between them. Finally, we conclude that the generating process of the oscillatory solution shown in [11] is as follows:

- For the system consisting of even inhibitory neurons, the solution generated by the first Hopf bifurcation (the UPS for \( k = 1 \)) is one-dimensionally unstable.

- When a number of coupled neurons is increased, stable dimension of the UPS for \( k = 1 \) is also increased.

- In a high-dimensional system, a solution from random initial states approaches the UPS and stays around it in a very long time. However, when a number of neurons \( n \) is small, this staying time is short and the solution converges to one of stable equilibrium points. When \( n \) is large enough, we cannot recognize that this oscillatory solution is a transient or steady state.

We confirm that this oscillatory solution is robust to injection of noise. The analysis of other connection types [16] such as small world network [17] is one of our open problems.

References


An Autoassociation Model
based on Entropy Functionals

Masahiro NAKAGAWA
Nagaoka University of Technology, Kamitomioka 1603-1, Nagaoka,
Niigata 940-2188, JAPAN
e-mail: masanaka@vos.nagaokaut.ac.jp

Abstract  In this paper, an entropy based associative memory model will be proposed and applied to memory retrievals with an orthogonal learning model to compare with the conventional autoassociative model with a quadratic Lyapunov functionals. In the present approach, the updating dynamics will be constructed on the basis of the entropy minimization strategy which may be reduced asymptotically to the autocorrelation dynamics as a special case. From numerical results, it will be found that the presently proposed novel approach realizes the larger memory capacity in comparison with the autocorrelation model based on dynamics such as association according to the higher-order correlation involved in the proposed dynamics.

1. Introduction
During the past quarter century, the numerous autoassociative models have been extensively investigated on the basis of the autocorrelation dynamics. Since the proposals of the retrieval models by Anderson,[1] Kohonen, [2] and Nakano,[3] some works related to such an autoassociation model of the inter-connected neurons through an autocorrelation matrix were theoretically analyzed by Amari, [4] Amit et al. [5] and Gardner.[6] So far it has been well appreciated that the storage capacity of the autocorrelation model, or the number of stored pattern vectors, \( L \), to be completely associated vs the number of neurons \( N \), which is called the relative storage capacity or loading rate and denoted as \( \alpha = L/N \), is estimated as \( \alpha \approx 0.14 \) at most for the autocorrelation learning model with the activation function as the signum one ( \( \text{sgn}(x) \) for the abbreviation).[7,8]

In contrast to the above-mentioned models with monotonous activation functions, the neuro-dynamics with a nonmonotonous mapping was recently proposed by Morita,[9] Yanai and Amari,[10] Shiino and Fukai.[11] They reported that the nonmonotonous mapping in a neuro-dynamics possesses a remarkable advantage in the storage capacity, \( \alpha \approx 0.27 \), superior than the conventional association models with monotonous mappings, e.g. the sigmoid or sigmoideal function.

In the present paper, we shall propose a novel approach based on the entropy defined in terms of the overlaps, which are defined by the inner products between the state vector and the embedded vectors.

2. Theory
Let us consider an associative model with the embedded binary vector e \( ^{(r)}\) \( _{i} \)=±1 (1\( \leq \)i\( \leq \)N,1\( \leq \)r\( \leq \)L), where N and L are the number of neurons and the number of embedded vectors. The states of the neural network are characterized in terms of the output vector \( s_{i} \) (1\( \leq \)i\( \leq \)N) and the internal states \( \sigma_{i} \) (1\( \leq \)i\( \leq \)N) which are related each other in terms of

\[
s_{i}=f(\sigma_{i}) \quad (1\leq i\leq N) .
\]

where \( f(\cdot) \) is the activation function of the neuron.

Then we introduce the following entropy which is to be related to the overlaps;

\[
I=\frac{1}{2}\sum_{r=1}^{L}m^{(r)^{2}}\log(m^{(r)^{2}}) ,
\]

where the overlaps \( m^{(r)} = \sum_{i=1}^{N}e^{(r)^{\dagger}}e_{i}^{(r)} \) are defined by

\[
m^{(r)}=\sum_{i=1}^{N}e^{(r)^{\dagger}}s_{i} ,
\]

here the covariant vector \( e^{(0)\dagger} \) is defined in terms of the following orthogonal relation,

\[
\sum_{r=1}^{N}e^{(r)^{\dagger}}e^{(s)}=\delta_{r,s} \quad (1\leq r,s\leq L) .
\]

\[
e_{i}^{(r)}=\sum_{r=1}^{L}a_{n}e_{i}^{(r)^{\dagger}} ,
\]

\[
a_{n}=\omega_{n}^{-1} ,
\]

\[
\omega_{n}=\left(\sum_{i=1}^{N}e_{i}^{(r)\dagger}e_{i}^{(r)}\right) .
\]

The entropy defined by eq.(2) can be minimized by the following condition

\[
\sum_{r=1}^{L}m^{(r)^{2}}=1 .
\]

(5b)
That is, regarding \([m^{(r)}]_1^{L} \) as the probability distribution in eq.(2), a target pattern may be retrieved by minimizing the entropy \(I\) with respect to \(m^{(r)}\) or the state vector \(s_i\) to achieve the retrieval of a target pattern in which the eqs.(5a) and (5b) are to be satisfied. Therefore the entropy function may be considered to be a functional to be minimized during the retrieval process of the auto-association model instead of the conventional quadratic energy functional, \(E\), i.e.

\[
E = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} s_i^\dagger s_j ,
\]

(6a)

where \(s_i^\dagger\) is the covariant vector defined by

\[
s_i^\dagger = \sum_{r=1}^{L} \sum_{j=1}^{N} e^{(r)}_{i} e^{(r)}_{j} s_j ,
\]

(6b)

and the connection matrix \(w_{ij}\) is defined in terms of

\[
w_{ij} = \sum_{r=1}^{L} e^{(r)}_{i} e^{(r)}_{j} .
\]

(6c)

According to the steepest descent approach in the discrete time model, the updating rule of the internal states \(\sigma_i\) \((1 \leq i \leq N)\) may be defined by

\[
\sigma_i(t+1) = -\eta \left( \frac{dI}{ds_i^\dagger} \right) ,
\]

(7)

where \(\eta (>0)\) is a coefficient. Substituting eqs.(2) and (3) into eq.(7) and noting the following relation with aid of eq.(6b),

\[
m^{(r)} = \sum_{i=1}^{N} e^{(r)}_{i} s_i = \sum_{i=1}^{N} e^{(r)}_{i} s_i^\dagger ,
\]

(8)

one may readily derive the following relation.

\[
\sigma_i(t+1) = -\eta \left( \frac{dI}{ds_i^\dagger} \right) = \eta \left( \frac{dI}{dm^{(r)}_{i}} \right) \sum_{j=1}^{N} e^{(r)}_{i} s_j^\dagger ,
\]

(9)

Generalizing somewhat the above dynamics, we propose the following dynamic rule for the internal states

\[
\sigma_i(t+1) = \eta \sum_{r=1}^{L} e^{(r)}_{i} \left[ \sum_{j=1}^{N} e^{(r)}_{j} s_j^\dagger \right] \left( \frac{1}{\alpha} + \alpha \log \left( \frac{1}{\alpha} + \alpha \right) \right) \frac{dI}{dm^{(r)}_{i}} ,
\]

(10)

In the limit of \(\alpha \to 0\), the above dynamics will be reduced to the autocorrelation dynamics.

\[
\sigma_i(t+1) = \eta \lim_{\alpha \to 0} \sum_{r=1}^{L} e^{(r)}_{i} \left[ \sum_{j=1}^{N} e^{(r)}_{j} s_j^\dagger \right] \left( \frac{1}{\alpha} + \alpha \log \left( \frac{1}{\alpha} + \alpha \right) \log \left( \sum_{j=1}^{N} e^{(r)}_{j} s_j^\dagger \right) \right) ,
\]

(11)

On the other hand, eq.(10) results in eq.(9) in the case of \(\alpha \to 1\). Therefore one may control the dynamics between the autocorrelation (\(\alpha \to 0\)) and the entropy based approach (\(\alpha \to 1\)).

3. Numerical Results

The embedded vectors are set to the binary random vectors as follows.

\[
e^{(r)}_{i} = \text{sgn} \left( z^{(r)}_{i} \right) \quad (1 \leq r \leq L) ,
\]

(12)

where \(z^{(r)}_{i}\) \((1 \leq i \leq N, 1 \leq r \leq L)\) are the zero-mean pseudo-random numbers between -1 and +1. For simplicity, the activation function, eq.(1), is set to

\[
s_i = f(\sigma_i) = \text{sgn}(\sigma_i) ,
\]

(13)

where \(\text{sgn}\) denotes the signum function defined by
\[ \text{sgn}(x) = \begin{cases} -1 & (x < 0) \\ 0 & (x = 0) \\ 1 & (x > 0) \end{cases} \]

The initial vector \( s_i(0) \) (1 ≤ i ≤ N) is set to

\[ s_i(0) = \begin{cases} e_{(s)}^{(i)} & (1 ≤ i ≤ H_d) \\ e_{(s)}^{(i)} & (H_d + 1 ≤ i ≤ N) \end{cases}, \]

where \( e_{(s)}^{(i)} \) is a target pattern to be retrieved and \( H_d \) is the Hamming distance between the initial vector \( s_i(0) \) and the target vector \( e_{(s)}^{(i)} \). The retrieval is succeeded if

\[ m^{(t)}(s) = \sum_{i=1}^{N} e_{(s)}^{(i)} s_i(t) \]

results in 1 for \( t ≥ 1 \), in which the system may be in a steady state such that

\[ s_i(t+1) = s_i(t) \]
\[ \sigma_i(t+1) = \sigma_i(t) \]

To see the retrieval ability of the present model, the success rate \( S_r \) is defined as the rate of the success for 1000 trials with the different embedded vector sets \( e_{(r)}^{(i)} \) (1 ≤ i ≤ N, 1 ≤ r ≤ L). To control from the autocorrelation dynamics after the initial state \( t \sim 1 \) to the entropy based dynamics \( t > T_{\text{max}} \), the parameter \( \alpha \) in eq.(10) was simply controlled by

\[ \alpha = \frac{T_{\text{max}}}{T_{\text{max}}} \alpha_{\text{max}} \]

where \( T_{\text{max}} \) and \( \alpha_{\text{max}} \) are the maximum values of the iterations of the updating according to eq.(10) and \( \alpha \), respectively.

Choosing \( N = 200 \), \( \eta = 1 \), \( T_{\text{max}} = 25 \), \( L/N = 0.5 \) and \( \alpha_{\text{max}} = 1 \), we first present an example of the dynamics of the overlaps in Figs.1(a) and (b) (Entropy based approach). Therein the cross symbols(×) and the open circles(○) represent the success of retrievals, in which eqs.(5a) and (5b) are satisfied, and the entropy defined by eq.(2), respectively, for a retrieval process. In addition the time dependence of the parameter \( \alpha/\alpha_{\text{max}} \) defined by eq.(18) are depicted as dots (.). In Figs.1 (a) and (b), after a transient state, it is confirmed that the complete association corresponding to eqs.(5a) and (5b) can be achieved.

Fig. 1 The time dependence of overlaps of the present entropy based model defined by eq.(10).

(a) \( H_d/N = 0.1 \)

(b) \( H_d/N = 0.3 \)

Fig. 2 The characteristics of the memory retrievals of autoassociation models.

(a) The dependence of the success rate on the loading rate \( \alpha = L/N \) of the present entropy based model defined by eq.(10).

(b) The dependence of the success rate on the loading rate \( \alpha = L/N \) of the association defined by eq.(11).
Then we shall present the dependence of the success rate $S_r$ on the loading rate $L/N$ are depicted in Figs.2(a) and (b) for $H_w=1$, $N=200$ for the entropy approach and the associatron, respectively. From this result, one may confirm the larger memory capacity of the presently proposed model defined by eq.(10) in comparison with the conventional autoassociation model defined by eq.(11). Comparing Figs. 2(a) and (b), it is found that the present approach may achieve twice of the memory capacity beyond the conventional autocorrelation strategy. Comparing Figs. 2(a) and (b), it is found that the present approach may achieve twice of the memory capacity beyond the conventional autocorrelation strategy.

4. Concluding Remarks

In the present paper, we have proposed an entropy based association model instead of the conventional autocorrelation dynamics. From numerical results, it was found that the large memory capacity may be achieved on the basis of the entropy approach. This advantage of the association property of the present model is considered to result from the fact such that the present dynamics to update the internal state eq.(10) assures that the entropy, eq.(2) is minimized under the conditions, eqs.(5a) and (5b), which correspond to the succeeded retrieval of a target pattern. In other words, the higher-order correlations in the presently proposed dynamics, eq.(10), which was ignored in the conventional approaches, [15] was found to play an important role to improve memory capacity, or the retrieval ability.

To conclude this work, we shall show the dependence of the storage capacity, which is defined as the area covered in terms of the success rate curves as shown in Fig.3, on the Hamming distance in Fig.3 for the autocorrelation learning model(CL) with eqs.(10) and $w_{ij}=(1/N)\Sigma e^{(i)}_j e^{(j)}$, instead of eq.(6c) as well as the orthogonal learning model(OL) with eqs.(10) and (6c). Therein one may see again the great advantage of the present model based on the entropy functional to be minimized beyond the conventional quadratic form [12,13]. In fact one may realize the considerably larger storage capacity in the present model in comparison with the associatron over $H_w/N=0.5$. The memory retrievals for the associatron become troublesome near $H_w/N=0.5$ as seen in Fig.3 since the directional cosine between the initial vector and a target pattern eventually vanishes therein. Remarkably, even in such a case, the present model attains a remarkably large memory capacity because of the higher-order correlations involved in eq.(10) as expected from Figs.1 and 2.

As a future problem, it seems to be worthwhile to introduce a chaotic dynamics in the present model introducing a periodic activation function such as sinusoidal one as a nonmonotonic activation function. [14] The entropy based approach[15] with chaos dynamics[14] is now in progress and will be reported elsewhere in the near future.

References

Analysis of Phase Noise and the Effect of Large Signal Nonlinearity in Dual Cross-Coupled LC Oscillators

Sohrab Samadian

Electrical Engineering Department, University of California, Los Angeles (UCLA), Los Angeles, CA 90095

Email: samadian@ucla.edu

Abstract-- Based on a physical process of conversion of device noises to output phase noise, an analysis is presented that explains the superiority of complementary oscillators in achieving higher figures of merit. Nonlinearity in the large signal operation of the cross coupled pair and its effect on phase noise degradation at higher amplitudes is quantified. Numerical findings are compared against measured data.

1. Introduction

Voltage controlled oscillators (VCOs) are continually the topic of active research. With increasingly more stringent requirements on their phase noise specifications, more efficient VCOs with higher figures of merits are needed in emerging bandwidth efficient standards. Different topologies have been compared already in a quest for the optimal performance. It has been observed that the complementary LC VCO of Fig. 1(b) can achieve a Figure of Merit (FOM) of up to 6 dB better than its NMOS- or PMOS-only counterpart (Fig. 1(a)) making it the topology of choice in power aware applications. To date, the most comprehensive attempts at the analysis of phase noise in complementary VCOs have solely used the LTV model developed by Hajimiri and Lee [1-4]. The reliance on this model alone and the ISF functions to calculate noise up-conversions can deprive circuit designers of gaining enough intuition needed in their design process. In turn, an analysis that physically explains the operation of the VCO with awareness of the nonlinear operation of the active cross coupled pair can lead to further circuit techniques to improve phase noise.

In this paper, we present a physical mechanism that explains the conversion of device noises to output phase noise in complementary LC oscillators. We will then compare the findings with the performance of N- or PMOS only oscillator. Increased phase noise at higher output amplitudes will be explained and quantified. Compared to the previous treatments of complementary oscillators [2-4], we believe this analysis is more comprehensive, simpler to understand and presents insights into the design of low noise VCOs.

2. Phase Noise Analysis

It has been observed that a complementary VCO can nominally generate an output amplitude two times larger than an NMOS VCO for the same tail current and the same equivalent parasitic resistance across the tank [3]. Dual cross coupled pair configuration in complementary VCO of Fig. 1(b) commutate the tail current as a square wave into the tank, the first harmonic of which defines the differential output amplitude: \( V_o = (4/\pi)IR \) where \( R \) is the differential equivalent resistance across the tank as depicted in Fig. 1(c). The underlying phase noise analysis converts the noise originating from any source to a single equivalent noise current across the tank shown in Fig. 1(c). Any such noise causes a differential voltage noise shaped by the tank’s transfer function:

\[
v_n = \frac{i_n \cdot R \left( \frac{\omega_n}{\omega_m} \right)}{2Q}
\]  

We will subsequently address each noise source separately.

2.1. Tank Noise

The parasitic resistance due to limited quality factor of the tank generates a stationary white noise that continuously injects a constant-PSD noise into the tank. However, time variant operation of the oscillator affects the conversion of this voltage noise to phase noise as quantified by the impulse sensitivity function (ISF) [1]. For an oscillator with a sinusoidal output voltage, the ISF is known to be \( \Gamma(\phi) = \cos(\phi) \), hence \( \Gamma_{\text{rms}}^2 = 1/2 \). The RMS value of the ISF function determines what percentage of the cyclostationary noise power converts to phase noise. Using (1), normalizing the noise power to differential signal power and multiplying by \( \Gamma_{\text{rms}}^2 \), the effect of tank noise on phase noise is:

\[
L_R(\omega_m) = \frac{4kTR^2}{V_o^2} \left( \frac{\omega_m}{2Q\omega_m} \right)^2
\]  

2.2. Noise in NMOS Cross-Coupled pair

Similar to the operation of the cross coupled pair in NMOS only VCO analyzed in [5], the NMOS cross coupled pair in Fig. 1(b) injects noise to the output almost entirely during the time window of zero crossing transition at the differential output voltage. Therefore a noise voltage \( v_n \) at the gate of transistors in cross coupled pair simply advances or retards zero crossing times, which in turn translates to phase noise. This can be modeled by a train of sampling pulses at two times the frequency of oscillation which sample the gate noise voltage to form output noise. In a simple model, assuming that the pair simply switches the differential current from -I to +I and that the output voltage is sinusoidal, it can be shown that the sampling pulse width is
Multiplication of these pulses with gate noise voltage translates to convolution of their spectra in frequency. Therefore, all the odd harmonics of the gate voltage noise are converted in frequency to the sideband PM noise at the output main tone. The effective equivalent stationary current noise power across the tank can be shown to be:

$$\frac{I_{nc}^2}{2} = 8kTf_n\pi V_o$$

Using (1) and (3), normalizing noise to the signal power, and replacing the amplitude of oscillation, we get:

$$L_c(\omega_m) = \frac{4kT\sqrt{V_o}}{2} \left( \frac{\omega_m}{2Q\omega_m} \right)^2$$

2.3. Tail Current Noise

The conversion of tail current noise to phase noise depends on its operation region. Moreover, circuit techniques can be used to filter out the critical noise around the second harmonic of the oscillation frequency [5]. At higher amplitudes of oscillation, tail current FET may operate in triode region. In this scenario, the small output resistance of the triode FET shunts most of the tail channel noise to AC ground. Therefore only a small fraction of the tail noise contributes to the output phase noise.

However, when the tail current FET operates in saturation region, its channel noise affects the output phase noise in a mixing action with the cross coupled pair. As explained in [5], the current commutation in cross coupled pair acts as a mixer that converts all the even harmonics (except low frequency) of tail current noise to close in output phase noise. A noise bin around 2kf_c is mixed down by the oscillator harmonics at (2k+1)f_c and (2k-1)f_c to form a correlated voltage output noise around the oscillation frequency f_o. The sum of all the frequency converted components of noise can be modeled by a stationary current noise across the tank:

$$\frac{I_{nt}^2}{2} = 4kTf_n\frac{2I}{V_{eff,t}}$$

where square law relationship is used to define the transconductance of the tail current FET; \( g_{mt} = 2I/V_{eff,t} \). Using (5) and (1), the resulting phase noise is:

$$L(\omega_m) = \frac{4kT\sqrt{V_o}}{2} \left( \frac{\omega_m}{2V_{eff,t}} \right)^2$$

2.4. Noise in PMOS Cross-Coupled Pair

The analysis of noise in the PMOS cross-coupled pair needs a different treatment. Noise current in lossy tank is stationary. The assumption of stationarity can also be applied to the tail current noise provided that the oscillation amplitude is not too large to push current source into triode at the peaks of oscillation. Simplification of stationarity of the cross coupled noise during a small window around zero crossing does not lead to significant error in calculated output noise. However, channel noise in PMOS cross coupled devices is injected directly to each side of the tank continuously and is cyclostationary (Fig. 2(a)). The question is how the single ended current noise source as shown in Fig. 2(a) affects the differential output phase noise. The single ended current noise can be equivalently represented by three noise currents as depicted in Fig. 2(b). Only half of the noise current (1/4th the noise power) represented by the noise source across the tank converts to phase noise. The rest of the noise converts to AM noise which can be removed by subsequent limiters. The conversion of this cyclostationary differential noise current to the output phase noise depends on the amplitude of oscillation.

To address the cyclostationarity of this noise, one must consider the effective ISF [1], defined by:

$$\Gamma_{eff}(\phi) = \Gamma(\phi) \cdot \alpha(\phi) \cdot G(t)$$

where \( \alpha(\phi) \) is the cyclostationarity factor. It turns out that with an assumption of sinusoidal voltage across the tank and the square law relationship of current, \( \alpha(\phi) = \beta(\phi) \).

$$\alpha(\phi) = \beta(\phi) \left[ 1 + m \sin(\phi) \right]$$

where \( m = \frac{\omega_m}{2V_{eff,p}} \).

$$\beta(\phi) = \begin{cases} 1 & 0 < \phi \leq \pi \\ 0 & \pi < \phi \leq 2\pi \end{cases}$$

When the oscillation amplitude is small compared to \( 2V_{eff,p} \), m < 1. Thus the effective ISF can be approximated as:

$$\Gamma_{eff, rms, small} = \frac{1}{2} \left( 1 + \frac{m^2}{4} \right)^{1/2}$$

Therefore when the oscillation amplitude is small, half of the differential noise converts to phase noise according to (10). Subsequently, the overall contribution of PFET pair to phase noise reduces to:

$$L_{pmos}(\omega_m) = 4kT\frac{\sqrt{V_o}}{2} \left( \frac{\omega_m}{2Q\omega_m} \right)^2$$

where \( I_{mp} = 4kTg_{mpo}R \) has been used for the thermal channel noise of PFETs at equilibrium bias point and \( g_{mpo,R} \), the overdrive gain of PFETs at equilibrium, is assumed approximately 1 for small amplitudes. To maintain higher oscillation amplitudes, the overdrive gain of the cross coupled pairs must be increased above 1. Fig. 3(b) shows the simulated large signal behavior of the PMOS and NMOS cross coupled pair in a complementary VCO with large out-
put voltage sweep. The hysteresis is due to the delay in cross coupled pairs at a high oscillation frequency (here $f_o=1.2\text{GHz}$). When the VCO operates in the voltage-limited regime, the cross coupled pairs enter their non-linear regions for a large part of the oscillation period. In steady state the effective transconductance of the PMOS cross coupled pair ($G_{mp}$) and that of the NMOS pair ($G_{mn}$) must cancel parasitic resistance of the tank such that $G_{mp} + G_{mn} = 2/R$. With equal contributions of negative resistance from the NMOS and PMOS cross coupled pair, $G_{mp} = G_{mn} = 1/R$. From the nonlinear curves in Fig. 3 it can be observed that for this condition to hold at large output swings when the PMOS pair exhibits high nonlinearity, the overdrive gain $g_{mp}R$ must be greater than 1; often at least 2 or 3. Using (8) for large amplitudes, the phase noise due to PMOS cross coupled pair at high oscillation amplitudes is approximately:

$$L_{pmos}(\omega_m) = \frac{4kTfR}{V_o^2} \left( \frac{V_o}{2Q\omega_m} \right)^2$$

(12)

where square law relationship has been used for transconductance of PMOS FETs.

### 3. Discussions and Design Implications

From the preceding noise analysis, the overall phase noise can be modeled as:

$$L(\omega_m) = \frac{4kTfR}{V_o^2} \left( \frac{V_o}{2Q\omega_m} \right)^2$$

(13)

where the noise factor $F$ is:

$$F = 1 + \gamma_n \frac{V_o}{2} + \gamma_n \left( \frac{\pi V_o}{4V_{eff,p}} \right)$$

(14)

when the oscillation amplitude is small enough to prevent PMOS cross coupled pair to experience high nonlinearity. Otherwise the noise contribution factor is approximately:

$$F = 1 + \gamma_n \left( \frac{m^2 V_o}{8} \right) + \gamma_n \left( \frac{\pi V_o}{4V_{eff,p}} \right) + \gamma_n \left( \frac{\pi V_o}{2V_{eff,d}} \right)$$

(15)

Noise filtering technique [6] and the operation of VCO’s tail current device in triode region can reduce the last factor in (14) and (15). For comparison, the noise factor of the NMOS only LC oscillator of Fig. 1(a) is:

$$F = 1 + \gamma_n \left( \frac{\pi V_o}{2V_{eff}} \right)$$

(16)

From (14) and (16) it can be concluded that the contribution of each cross coupled pair to the output phase noise in the complementary VCO is half of the NMOS pair in Fig. 1(a). Therefore with $\gamma_n \approx \gamma_p$, a 6 dB improvement in phase noise may be expected. However, as the oscillation amplitude increases further, the PMOS pair contributes more to the phase noise as predicted in [2] and quantified in this paper in (15). A closer look at (15) suggests that the contribution of PMOS cross coupled pair can be reduced at higher amplitudes by reducing their equilibrium transconductance $g_{mp}$. To illustrate this, two complementary LC VCOs were designed in a standard 0.18 $\mu$m CMOS process, shown in Fig. 4. Noise filtering technique was used to reduce the effect of tail FET channel noise to negligible values. The two VCOs use the same tank, oscillate at about the same frequency and consume almost equal powers. In OSC1, PMOS and NMOS are sized to deliver equal transconductances, as commonly used. However, the sizings in OSC2 are chosen to reduce $g_{mp}$. In both VCOs, $g_{mp}+g_{mn}$ are about the same. The simulated phase noise at 10MHz away from the carrier in OSC2 is 2.1 dB lower than OSC1, proving the preceding analysis. The choice of unequal values of transconductance can result in asymmetry of the waveforms leading to increased low frequency phase noise [1]. However, we did not observe such effect in SpectreRF simulations for the VCO of Fig. 4.

### 4. Experimental Verification

The validity of phase noise expressions in section II has been verified with numerous simulations in SpectreRF. Kao et. al. present a qualitative analysis of a complementary VCO which does not offer a closed form equation for the excess noise factor $A$ [4]. We received the circuit schematics from Kao et. al. and simulated the VCO using the fabricated technology model files. The VCO was implemented in 0.35 $\mu$m CMOS technology. We verified all the simulated operating parameters of the VCO (frequency, amplitude, current) against the reported values in [4]. The PMOS cross coupled pair partly operates in the nonlinear region of operation. We calculated all the parameters including $\gamma_n = 0.79$ and $\gamma_p = 0.84$ from model files and simulations. Factor $m$ in (8) is 2.93. Eq. (15) predicts a noise factor $F = 8.89$. The equivalent noise factor in [4] is defined as $2(A+1) = 7.02$; slightly lower than the more rigorous calculation using Eq. (15). However, [4] reports $A$ from simulations only. Fig. 5 shows a comparison between the measured phase noise reported in [4] and the calculated phase noise from (13) and (15). The VCO oscillates at close to 2GHz. The calculated phase noise at 600kHz away from the carrier is -119.4 dBc/Hz; close to the measured value of -116 dBc/Hz. Flicker noise effects dominate at frequency offsets below 100KHz, where the measured and predicted graphs of phase noise start to depart.

### 5. Conclusion

An analysis of the phase noise in complementary LC oscillators based on the physical process of device noise to phase noise conversion was presented and compared with VCOs with single cross coupled pair. It was proved that while at low oscillation amplitudes, when the auxiliary cross coupled pair exhibits less nonlinearity than it would at high swings, the complementary VCO provides an FOM of up to...
6 dB higher than its single cross coupled counterpart. In addition, a major difference between the impact of the PMOS device noise and that of the NMOS pair in Fig. 1(b) is that the former injects cyclostationary noise across the tank continuously while the latter introduces noise only during output voltage transitions. Thus the high impedance of the tail current is beneficial to reducing phase noise contribution of the NMOS devices. A measured data by Kao et al. was compared to the analysis for validation.

References


Fig. 3. Large signal characteristic of the cross coupled pairs (Vdd=1.8V)

Fig. 4. Reduction of thermal induced phase noise by choosing unequal transconductances in NMOS and PMOS cross-coupled pairs.

Fig. 5. Measured Phase noise of the VCO in [4] compared with predicted curve from the analysis ((13) and (14)).
3 GHz Spread Spectrum Clock Generator for Serial ATA-II using Random Frequency Modulation

Luca Antonio De Michele†, Fabio Pareschi†., Riccardo Rovatti†,# and Gianluca Setti†.†

† ARCES, University of Bologna, via Toffano, 2/2, – Bologna 40125, ITALY
‡ ENDIF, University of Ferrara, via Saragat, 1, – Ferrara 44100, ITALY
# DEIS, University of Bologna, vialle Risorgimento, 2 – Bologna 40135, ITALY

Email: ldemichele@arces.unibo.it, fparesci@arces.unibo.it, rrovatti@arces.unibo.it, gsetti@arces.unibo.it

Abstract—This paper proposes a Spread Spectrum Clock Generator specifically meant for 3 GHz Serial ATA-II applications. The spreading is obtained through a PLL frequency modulation driven by a random signal, to avoid strict periodicity of the clock and thus high peaks in its power spectral density. The circuit, including a PLL-modulator and a Random Bit Generator has been designed in UMC 0.18 μm technology; the random bit generator is based on a chaotic-map. Simulations show Electromagnetic Compatibility improvement of approximately 13 dB with respect to the reference case. A prototype of this circuit is currently under fabrication.

1. Introduction

Reduction of Electro-Magnetic Interference (EMI) in electronic devices is an issue of increasing interest for designers. Preeminent interference sources are synchronization signals, due to their sharp edges and their periodic nature, which concentrate power at multiples of their frequency; in particular for high-speed interfaces between signal processing units and the other peripheral storages, like HDD and DVD/CD.

To cope with all problems related to the increasing data-rate of these peripheral devices, the Serial ATA (Advanced Technology Attachment) protocol [1] has been introduced and it is now imposing as the new standard de facto. This computer bus technology, primarily designed for fast data transmission to and from Hard Disk Drives, has been designed to transmit data through a simple serial cable, in opposition to the previous ATA standard that relies on parallel data transmission. In this way, all transmission related problems are now concentrated into a single high-speed line, making the system simpler and more reliable. Recently this protocol has raised its clock rate to 3.0 GHz, with the release of SATA–II.

The obvious drawback is related to EMI increase, due to the higher transmission rate. To this regard, Serial ATA specifications formaly mention the possibility of introducing spread spectrum techniques on its synchronization signal, even stating its allowed amount, thus implicitly suggesting that a spread spectrum clock can conveniently be applied to perform an on-chip EMI reduction, without any need of heavy shielding materials and so reducing the overall cost of the equipment.

Spread Spectrum clock consists of a frequency modulation producing a clock signal whose edges are slightly delayed or anticipated avoiding perfect periodicity. In this way, all peaks in the power spectrum are sensibly reduced. This point of view is perfectly coherent with FCC and CE regulations [2] that link compliance with the ability of fitting the interfering power spectrum within a prescribed mask.

Of course, in this way the synchronization process results impaired; for the proper operation of the system it is necessary that all connected devices are designed to be compatible with the introduced modulation. SATA specifications set the maximum allowed value of the frequency deviation to 0.5% of the fundamental frequency.

On chip – or design-time – solutions [3], assure that the implemented electronic equipment generates electromagnetic interference with power spectral density as flat as possible, so that its integral within any sufficiently narrow frequency range (and therefore in the bandwidth of any unintentional receiver) is as low as possible.

It can be intuitively accepted that the efficiency of these methods critically depends on the statistical properties of the modulating signal. Signals produced by random systems seem particularly appealing from this point of view, since they are typically aperiodic and broadband. Random-based frequency modulation [4, 5] has been shown to significantly reduce the EMI due to clock signals with respect to results achieved in classical literature [6, 7]. In those papers, emphasis is on continuous-valued random modulation with large modulation indexes (slow-modulation) for which an analytical estimation of the spectrum profile can be provided.

More recently, in order to solve known problems connected with slow modulation, it has been introduced in [8] a binary fast random modulation, which shows even better flattening properties with respect to slow modulation, as long as it operates at proper modulation indexes, derived by means of numerical optimization.

In this paper we propose a Spread-Spectrum Clock Generator (SSCG) designed to implement a fast binary modulation. The SSCG architecture is based on a PLL with few modifications to achieve a frequency modulator, which is designed to work with a mean frequency \( f_0 = 3 \) GHz. A random bit generator, whose purpose is to generate the driving signal is also implemented; it is designed to work up to 45 MHz and it can be shown that its realization can conveniently be derived by already existing circuit blocks [9].

The circuit is currently being manufactured with a 1.8V...
0.18 µm CMOS single-poly six-metal process. Our implementation, as well as the achieved results, is described in the following.

2. Architecture of the Spread Spectrum Clock

The block diagram of the SSCG is shown in Fig. 1.

In the following we will introduce the adopted modulation and the implementation of the main blocks.

2.1. Binary Fast Random Frequency Modulation

Let us consider clock signal \( s(t) \) as the result of a frequency modulation:

\[
s(t) = \text{sgn} \left\{ \cos \left[ 2\pi \left( f_0 + \Delta f \int_{-\infty}^{t} \zeta(\tau)d\tau \right) \right] \right\}
\]

where \( f_0 \) indicates the carrier frequency, \( \Delta f \) the frequency deviation and \( \zeta(t) \) the driving PAM signal:

\[
\zeta(t) = \sum_{k=-\infty}^{\infty} x_k g(t-kT)
\]

given that \( g(t) \) is a unit pulse of duration \( T \) and that \( x_k \) are random values constituting the modulating sequence, which can only assume the values -1 and +1. This is known as binary modulation.

It has been shown [8] that this modulation flattens the first harmonic in the PSD when the modulation index \( m = \Delta f \cdot T \) is equal to \( m_{\text{opt}} \approx 0.318 \). Furthermore, the observed reduction of the peak amplitude is the best with respect to all other known modulations.

2.2. Chaos-based Random Bit Generator (RBG)

The purpose of the RBG is to generate the PAM driving signal. Since the performances of the proposed binary modulation strongly depend on symbols being uncorrelated, the choice of the RBG is critical. Furthermore, being a fast modulation, the driving signal is required to work at high frequency, i.e. the random bit generator must achieve a high throughput. In other terms, pseudo-random generators based on linear feedback shift registers, which are capable of achieving a very high throughput, produce sequences that can be far from optimum for this application, depending on the length of the register. Also, classical true-random number generators based on the observation of physical random phenomena such as thermal noise, which achieve the best results in terms of signal low correlation, are not suitable for this application, due to their very low throughput. Hence, to achieve our goal we follow [9] and base our design on a chaotic map [10]. With this solution we are able to obtain a throughput up to about 45 Mbit/s, which is the speed necessary to achieve the optimum modulation index at the maximum frequency deviation allowed by the SATA protocol.

2.3. PLL-based Frequency Modulator

The block diagram of the modulator is shown in Fig. 2. The core block of the PLL in Fig. 3-(a) is the same as a conventional PLL-based clock generator. It includes a reference clock oscillator, a Phase-Frequency Detector (PFD), a Charge Pump (CP), a Low-Pass Filter (LPF), a Voltage-Controlled Oscillator (VCO) and a Divider by \( N \) on the feedback path. Its purpose is to set the output frequency \( f_{\text{out}} = N f_0 \), where \( f_0 \) is the frequency of the reference clock. From standard PLL analysis [11, 12] we know that its closed-loop transfer function \( f_{\text{out}}(s)/f_0(s) \) has a low-pass nature, with cut-off frequency \( \omega_c \).

The conventional scheme is indeed modified with the addition of a driving signal between LPF output and VCO input. If we suppose that this signal is high frequency with respect to \( \omega_c \), we can notice that it drives the VCO as in an open-loop system, since it cannot pass through the loop composed by the divider, the PFD, the CP and the LPF, due to the low-pass nature of the loop. [13]

With these assumptions, the PLL can be considered as a frequency modulator, where the modulating signal is the above driving signal, and the carrier \( f_0 \) is the output frequency \( f_{\text{out}} \) of the PLL working without any modulation.

The core block of the PLL is the VCO in Fig. 3-(a). Essentially it consists of a Ring–Oscillator (Fig. 3-(c)), whose basic stage is modified (Fig. 3-(d)) in order to minimize the power consumption while maximizing the oscillation frequency [14]. The classic common path for the inputs of the inverter has been split in two: the input for the NMOS transistor comes from the preceding inverter, whereas the input for the PMOS transistor comes from the inverter two positions before, to have an anticipated signal with the same phase with respect to NMOS input. In this way the PMOS, which is slower than the NMOS, is turned on in advance, allowing faster operations.

An additional PMOS transistor, driven by the control voltage \( V_{\text{ctrl}} \), realizes the modulation in the VCO, controlling the current flowing through the inverters.

The additive function in the modulation process is instead performed by the IS (Fig. 3-(b)), through two pass-transistors driven by \( \Phi_1 \) and \( \Phi_2 \), along with the two cur-
Figure 3: (a) Block diagram of the VCO; (b) Input Stage; (c) Ring-Oscillator; and (d) Ring Oscillator basic stage.

Figure 4: Voltage/Frequency characteristic of the VCO in non spread spectrum mode (solid line) and spread spectrum mode (dashed lines).

The circuit has been completely designed in UMC 0.18µm CMOS technology. The layout of the designed circuit is shown in Fig. 5. Some post-layout simulations, which include parasitic effects and confirm the effectiveness of the proposed implementation are presented. The center-spread frequency has been set to the nominal value of \( f_0 = 3 \) GHz and the chaotic map bit-rate to \( f_m = 47.17 \) Mbit/s. The frequency deviation \( \Delta f \) is set to \( \Delta f = 15 \) MHz to achieve the optimal modulation index \( m \) value.

Fig. 6-(a) represents the power spectrum density of the output clock signal. The simulated spectrum is also compared with the theoretical one from [8]. As can be seen, the simulated spectrum is very close to the theoretical one. Fig. 6-(b) shows a comparison between the simulated power spectrum density of the output clock signal and the same spectrum obtained from the circuit without any driving signal, i.e. working as a standard PLL-based clock generator. The resolution bandwidth is set to 120 kHz, as
Figure 6: (a) Comparison between power spectrum density of the output clock obtained from the simulated circuit and the theoretical power spectrum density of the binary modulation; and (b) comparison between power spectra density of the modulated and non modulated output clock. The spectra are measured in $\text{dBV}^2$, with RBW = 120KHz.

indicated by CISPR regulations [2][15]. The comparison shows a peak reduction on the fundamental tone of about 13 dB.

4. Conclusion

In this paper, a spread-spectrum clock generator for SATA-II applications implementing a binary frequency modulator is presented. This circuit is driven by a chaotic map, which implements a fast true-random bit generator. The circuit has been designed in a commercial CMOS technology and simulations have been presented, which confirm that the proposed circuit works effectively as a binary frequency modulator, as expected from the circuit model; furthermore, peak amplitudes in the power spectrum are attenuated and the proposed architecture does achieve the spread-spectrum function as expected.

References

Mathematical Analysis of Injection-Locked Frequency Dividers

Tao Xu∗, Zhipeng Ye† and Michael Peter Kennedy‡

∗†‡Department of Microelectronic Engineering, University College Cork, Ireland
‡Tyndall National Institute, Lee Maltings, Cork, Ireland
Email: taox@ue.ucc.ie, z.ye@mars.ucc.ie, peter.kennedy@ucc.ie

Abstract—The injection-locked frequency divider (ILFD) is investigated. A novel model is derived and it is demonstrated that the ILFD works as a phase-locked-loop (PLL) when locked. The qualitative reasons why the frequency is divided and why it is mainly locked at even numbers are introduced. The theory is validated by experimental results.

1. Introduction

Phase-locked loops (PLL) are widely used in wireless communication applications such as frequency synthesis. The frequency prescaler is most used in the feedback of a frequency synthesizer in order to divide the frequency by a fixed number. Frequency dividers are categorized into two groups, i.e. injection-locked (ILFD) and static [1] frequency dividers. Static dividers offer both wide-bandwidth and high power consumption. On the contrary, ILFDs consume lower power at the expense of a narrow locking range.

In this paper we analyse an LC oscillator based injection-locked frequency divider from a mathematical respective. Until now, almost all the researchers considered the ILFD as just an oscillator with a special purpose [2], [3]; we point out that it works as a type of PLL when a signal is injected. A novel model is introduced in Section 2 to help us to understand it more clearly. In Section 3, we compare the reasons and conditions for locking of the ILFD with that of the PLL. We show why this type of frequency divider favors specific division ratios as well. Experimental results are shown in Section 4.

2. An Accurate Model of ILFD

The circuit schematic of the injection-locked frequency divider we consider has been described in [4], [5], and is shown in Fig. 1. It can be simplified to the ideal circuit shown in Fig. 2. \( N_R \) represents the non-linear resistor composed of the components in Fig. 1 other than \( R_s \), L and C.

When there is no injected signal, we can get the expression of this topology as below

\[
C \frac{dV_C}{dt} = I_L - f(V_C) \\
L \frac{dI_L}{dt} = -I_L R - V_C,
\]

where \( f(V_C) \) is the driving point characteristic of the non-linear resistor which can be obtained by both experiment and Spice simulation. The driving-point characteristics derived from Spice simulations and ex-
periments are shown, together with a cubic approximation, in Fig. 3. The cubic approximation gives:

\[ f(V_C) = aV_C - \frac{a}{V_{DD}^2} V_C^3, \]  

(2)

where parameter \( a \) is the function of the bias voltage of the tail transistor \( M_6 \).

This free-running oscillator can be viewed as a feedback system. The model of it is shown in Fig. 4.

![Figure 3: Driving point characteristics: Solid: Experimental; Dotted: Cubic Fit To Experiment; Dashed: PSpice](image)

Figure 3: Driving point characteristics: Solid: Experimental; Dotted: Cubic Fit To Experiment; Dashed: PSpice

Now substituting (3) into (2), we obtain the governing equations for the ILFD:

\[ C \frac{dv_C}{dt} = I_L - (A + daV_{in})V_C + \frac{A + daV_{in}}{V_{DD}^2} V_C^3. \]  

(4)

\[ L \frac{di_L}{dt} = -I_L R - V_C, \]  

(5)

where \( A \) and \( da \) are used to represent \( \hat{a}(V_{GS}) \) and \( \frac{da}{dV_{GS}} \) in (3).

Equation (4) is modified, by collecting the factors \( A \) and \( daV_{in} \), to obtain

\[ C \frac{dv_C}{dt} = I_L - A(V_C - \frac{V_C^3}{V_{DD}^2}) - daV_{in}(V_C - \frac{V_C^3}{V_{DD}^2}). \]  

(6)

The difference between (1) and (6) is \( -daV_{in}(V_C - \frac{V_C^3}{V_{DD}^2}) \). Thus, a novel model is derived, as shown in Fig. 5.

![Figure 4: Model of the free-running oscillator](image)

Figure 4: Model of the free-running oscillator

When a signal is injected into the circuit shown in Fig. 1, the parameter \( a \) is modified to

\[ a = \hat{a}(V_{GS} + v_{GS}) \approx \hat{a}(V_{GS}) + \frac{d\hat{a}}{dV_{GS}} \Big|_{V_{GS}} v_{GS}, \]  

(3)

where \( \hat{a} \) is a function of both \( V_{GS} \) and \( v_{GS} \).

As can be seen from Fig. 5, the injected signal goes through a multiplier first, then it meets the RLC tank after combining with the signal in the interior feedback loop of the free-running oscillator. Since an analog multiplier acts as a phase detector [6], it can be concluded that the injection-locked frequency divider is equivalent to a dual-loop phase-locked-loop (PLL) at the mathematical level.

The RLC tank acts as a band-pass filter to reject the frequency components far from \( \omega_0 \) [7]. It can be expressed as

\[ H(j\omega) = \frac{R + j\omega L}{1 + j\omega RC + (j\omega)^2 LC}. \]  

(7)

The two negative feedback loops separate after the nonlinear device, and each one has an individual gain \( A \) and \( da \). The output of the nonlinear device \( v_{nd}\), shown in Fig. 5, is obtained as below

\[ v_{nd} = 1 - \frac{v_{o}^2}{V_{DD}^2}, \]  

(8)
Because of the effect of the nonlinear device, the injected voltage signal $V_{in}$ is transferred to a current signal $I_{inj}$ to affect the free-running oscillator after multiplying with the signal in the “da feedback loop”. This model is more accurate and clear than the previous work [8], [9].

3. Mathematical Analysis

3.1. Analysis of PLL

The model of a typical PLL is shown in Fig. 6. This PLL example uses an analog multiplier as the phase detector in order to be more easily compared with the ILFD. $v_{pd}$ and $v_{lpf}$ are set to be the outputs of the phase detector and the low-pass filter respectively. We assume

$$v_{in} = V_{in}\sin(\omega_{in}t + \varphi), \quad (9)$$

$$v_{o} = V_{o}\sin(\omega_{o}t), \quad (10)$$

where $V_{in}$ and $V_{o}$ are the amplitudes of $v_{in}$ and $v_{o}$, $\omega_{in}$ and $\omega_{o}$ are frequencies and $\varphi$ is the phase angle between input and output signals. In this case,

$$v_{pd} = v_{in} \times v_{o} = \frac{V_{in}V_{o}}{2}\cos[(\omega_{in} - \omega_{o})t + \varphi] - \frac{V_{in}V_{o}}{2}\cos[(\omega_{in} + \omega_{o})t + \varphi]. \quad (11)$$

The low-pass filter rejects the higher frequency $(\omega_{in} + \omega_{o})t + \varphi$. Then

$$v_{lpf} = \frac{V_{in}V_{o}}{2}\cos[(\omega_{in} - \omega_{o})t + \varphi]. \quad (12)$$

The PLL is locked when $v_{lpf}$ is constant, i.e. $\omega_{in} = \omega_{o}$.

![Figure 6: Model of PLL](image)

3.2. Analysis of ILFD

The ILFD is similar to the PLL except that in this case the locking depends on the output of the phase detector $I_{inj}$. Firstly, we set the same input and output signals as those in the PLL, i.e. (9), (10). As shown in Fig. 5, it’s relatively straightforward to calculate the output of the phase detector:

$$I_{inj} = v_{o} \cdot V_{o} \cdot da \cdot v_{in} = \frac{1}{2} \frac{1}{4} (V_{o} - 3 V_{o}^{3}) \cos[(\omega_{in} - \omega_{o})t + \varphi] - \frac{1}{2} \frac{1}{4} (V_{o} - 3 V_{o}^{3}) \cos[(\omega_{in} + \omega_{o})t + \varphi] + \frac{1}{8} \frac{1}{8} \cos[(\omega_{in} - 3\omega_{o})t + \varphi] - \frac{1}{8} \frac{1}{8} \cos[(\omega_{in} + 3\omega_{o})t + \varphi]. \quad (13)$$

From (13) it is found that the phase of $I_{inj}$ may be $(\omega_{in} \pm n\omega_{o})t + \varphi$. The RLC tank filters out all frequencies other than $\omega_{o}$. Hence, the ILFD is locked only when $\omega_{in} \pm n\omega_{o} = \omega_{o}$, i.e. $\omega_{in} = (\pm n + 1)\omega_{o}$. Because the nonlinear resistor has a cubic approximate characteristic, $n$ is always an odd integer, so that the ratio of

$$\frac{\omega_{in}}{\omega_{o}} = \pm n + 1 \quad (14)$$

should be an even number. This explains qualitatively why we get strong lockings at even ratios of $\omega_{in}/\omega_{o}$, but not at odd multiplies of $\omega_{o}$. In this example, the frequency divider will be locked when $\omega_{in} = 2\omega_{o}$ or $\omega_{in} = 4\omega_{o}$.

Since the driving point characteristic is not pure cubic, there may be other values of $\omega_{in}$ that cause the ILFD to be locked. For instance, $\omega_{in} = \omega_{o}$ or $\omega_{in} = 1.5\omega_{o}$, but the locking ranges at these points are very small in comparison with those at even numbers. This is demonstrated in the experiments introduced in the next section.

3.3. Comparison

The ILFD and PLL have the qualitatively similar locking behaviour. However, they have two key differences. First, the ILFD uses the RLC tank to act as a band-pass filter while the PLL has a low-pass filter. This results in different locking conditions and allows an oscillator to become a frequency divider. Second, there is a non-linear device in the feedback loop of the ILFD, which makes the expression of $I_{inj}$ more complex than the corresponding $v_{pd}$ in a PLL. This induces the ILFD to be locked at a few different values of $\omega_{in}/\omega_{o}$.

4. Experimental Results

The locking range is most easily observed in a devil’s staircase diagram [10], [11], as shown in Fig. 7. From the staircase diagram, it is clear that there are strong lockings at division ratios of 2 and 4, as predicted in Section 3.
Figure 7: Experimentally measured Devil’s staircase diagram showing strong lockings at $\frac{\omega_{\text{inj}}}{\omega_o} = 2$ and 4.

On closer examination, note that the ILFD is also locked at odd division ratios and a few fractions, but their locking ranges are very small. As highlighted in Fig. 8, the ILFD is locked at $\frac{\omega_{\text{inj}}}{\omega_o} = 1, \frac{4}{3}$, and $\frac{3}{2}$. Closer examination shows further lockings in the sequence, albeit with smaller locking ranges, which are coincident with the “Farey Sequence” [12].

Figure 8: Experimentally measured Devil’s staircase diagram showing “Farey Sequence” of locking at $\frac{1}{1}$, $\frac{4}{3}$ and $\frac{3}{2}$.

5. Conclusion

We have introduced a novel model which suggests that the ILFD behaves like a dual-loop PLL at a mathematical level. Unlike the normal PLL, the RLC tank eliminates frequencies far from $\omega_o$, which allows the system to be used as a frequency divider. The cubic characteristic of the non-linear resistor is the reason why the ILFD is preferentially locked at even ratios of $\omega_{\text{inj}}/\omega_o$ and specially at $\frac{2\omega_{\text{inj}}}{\omega_o} = 2$ and $\frac{2\omega_{\text{inj}}}{\omega_o} = 4$. The existence of a larger “Farey Sequence” of lockings is confirmed. This needs to be investigated further.

Acknowledgments

This work is supported in part by Science Foundation Ireland under Grant 02/IN.1/145.

References

Analysis and Evaluation of Harmonic Distortion in the Hartley Oscillator

Gaetano Palumbo, Melita Pennisi, Salvatore Pennisi

University of Catania
DIEES - Dipartimento di Ingegneria Elettrica Elettronica e dei Sistemi
Viale Andrea Doria, 6, I-95125 CATANIA – ITALY
Phone:+39-095-7382313; Fax:+39-095-330793
e-mail: [gpalumbo, mpennisi, spennisi]@diees.unict.it

Abstract— In this communication a simple technique based on the phasor’s method is presented in order to analytically deal with nearly sinusoidal oscillator constraints. Unlike other methods, more complex from a computational point of view, such as Volterra series and Poincaré’s perturbation, we are able to set the main linearity performance of a weakly nonlinear oscillator through closed form relationships. This gives to the designer a deep insight into the circuit behavior. In particular, we applied the methodology to the case of the Hartley oscillator, finding the expression of amplitude and frequency of the fundamental, and second- and third-order harmonic distortion factors as functions of the circuit components. Simulations performed with Spectre are provided and the results are found in good agreement with those expected.

1. Introduction

Nearly sinusoidal oscillators are widely used in modern electronics, especially in communication systems [1-2]. Considering their typical application fields, such as the mixer circuits of RF communication systems, where the input signal must have an adequate linearity level, it is clear how this constraint is crucial for any oscillator designer. Yet practical oscillators are intrinsic nonlinear circuits, since the gain stage is required to be nonlinear in order to exhibit an amplitude stabilization mechanism. Consequently, it is useful to develop a method that allows the designer to simply cope with circuit components to analytically set the main performance of a nearly sinusoidal oscillator.

Several theoretical and experimental papers have been devoted to nearly sinusoidal oscillators [3-5]. The analysis developed in [3] uses Volterra kernels and allows the evaluation of oscillation frequency and amplitude of the fundamental, by solving two nonlinear equations generated by a recursive algorithm. It reveals a good agreement with experimental results, but only after a large number of iterations. This method however is not suitable for a pencil and paper design and no information about harmonic distortion is outlined.

Another analysis procedure is presented and exploited in [4-5], where Poincaré’s perturbation method for weakly nonlinear system is exploited. The core of this method consists of a differential equation (the so called determining equation) containing the small parameter $\varepsilon$. By using a recursive method it is possible to solve the determining equation and to get the expression of the signal components at fundamental and its multiple frequencies. However, this approach is still expensive in terms of computation time.

Recently, an efficient method to evaluate nonlinearities in feedback amplifiers has been presented and discussed in [6-8]. Here, closed-form expressions of second- and third-order harmonic distortion factors are derived by applying the phasor notation and the method’s accuracy was validated through several examples.

In this work, we extend the field of application of the above approach to design and analyze the Hartley oscillator. The approach is simple and amenable for a manual design. It is based on the separation of the nonlinear (static) function from the frequency dependent linear transfer function. Phasor notation is adopted and only simple algebraic manipulation is exploited. The goal of our approach is the formulation of a set of algebraic equations which allow the fundamental design parameters, which are the oscillation frequency, the amplitude of the fundamental harmonic and the second- or third-order harmonic distortion factors, to be established.

In particular, we develop in section 2 the analysis of the Hartley oscillator and derive a set of equations useful for designing it. In section 3 we apply the equations found to a design example and perform simulations to verify their accuracy. Finally, in section 4 some conclusions are given.
2. Hartley Oscillator

In the Hartley oscillator (shown in Fig. 1) the nonlinear device is the MOSFET, $M_1$, biased in its saturation region. In particular we suppose that oscillator behavior is weakly nonlinear. This means that we assume the transistor at the steady state never departs from its quiescent point.

Voltage source $V_{DD}$ gives a step voltage that biases the transistor and enables the oscillatory regime. Under suitable biasing, we can model the transistor at the steady state as a nonlinear voltage controlled current source as in Fig. 2, where the AC equivalent circuit is shown.

Let us now represent the time-domain steady state output voltage, $v_o(t)$, considering its first three harmonics only, through the phasor notation

$$v_o(t) = \text{Re}(V_1 e^{j\omega t} + V_2 e^{j2\omega t} + V_3 e^{j3\omega t})$$

where complex coefficients $V_i$ are in the form $|V_i| e^{j\phi_i}$, being $|V_i|$ the magnitude and $\phi_i$ the phase of the $i$-th harmonic. In particular, we can set $\phi_1=0$ and measure the phase shift of the second and third harmonics with respect to the first (therefore, this means that coefficient $V_0$ is real).

Consider the MOSFET drain current given by a third-order polynomial of the voltage across gate and source

$$i_{DS} = a_1 v_o + a_2 v_o^2 + a_3 v_o^3$$

where real coefficients $a_i$ are known constants with appropriate dimensions. Substituting (1) in (2), disregarding the DC component and all harmonics higher than the third order, and denoting as $V^*$ the complex conjugated of phasor $V$, we get the expression of $i_{DS}$ reported in (3) at the bottom of the page.

Collecting the terms with the same exponential, the expressions of $I_{DS}$ that appear in the latter of (3) are given by

$$I_{DS1} = \frac{a_1 V_1}{2} + \frac{3}{4} a_1 V_1^3$$

$$I_{DS2} = a_2 V_o + \frac{3}{2} a_2 V_o^2 + \frac{3}{2} a_2 V_1 V_o$$

$$I_{DS3} = a_3 V_o + a_3 V_o V_2 + \frac{a_1}{4} a_3 V_o^3$$

where, keeping in mind that, under the assumption of weakly nonlinear behavior, the magnitude of both $V_o$ and $V_o$ is much lower than $V_0$, we neglected the terms proportional to $V_o V_o$ and to $V_o V_o$ which appear in $I_{DS1}$, and high-order

$$I_{DS} = a_1 e^{j\omega t} + a_2 e^{j2\omega t} + a_3 e^{j3\omega t}$$
terms (products of \(V_{o2}\) and \(V_{o3}\)) that appear in \(I_{DS2}\) and \(I_{DS3}\).

Considering for simplicity \(L = L_1 = L_2\), we can compute the output signal as

\[
V_o = I_{DS} \cdot Z(j\omega)
\]

\[
= I_{DS} \cdot \frac{j\omega^3 L C}{1 - \omega^2 \left(2LC + \frac{L^2}{R^2}\right)} - j2\omega\frac{L}{R} (\omega^3 LC - 1)
\]

where the operator “\(\circ\)” means that the function at the right must be evaluated at the frequency of the incoming signal [8]. Hence the output signal is equal to

\[
v_o = \text{Re} \left( I_{DS} Z(j\omega_0) e^{j\omega_0 t} + I_{DS} Z(j2\omega_0) e^{j2\omega_0 t} + I_{DS} Z(j3\omega_0) e^{j3\omega_0 t} \right)
\]

Equating (6) and (1), since the coefficients of the exponentials with the same frequency must be equal, yields the following system in which we introduced the loop gain \(T(j\omega) = a_1 Z(j\omega)\)

\[
\begin{align*}
Z(j\omega_0) &= \frac{1}{a_1 + \frac{3}{4} a_2 V_{o2}^2} \\
V_{o2} &= \frac{1}{2 a_1} \frac{T(j2\omega_0)}{1 - \left(1 + \frac{3}{2} \frac{a_2}{a_1} V_{o2}^2 \right)^2} V_{o2}^2 \\
V_{o3} &= \frac{1}{4 a_1} \frac{T(j3\omega_0)}{1 - \left(1 + \frac{3}{2} \frac{a_2}{a_1} V_{o3}^2 \right)^2} V_{o3}^3
\end{align*}
\]

The set of equations in (7) allows us to set independently the oscillation frequency, \(\omega_0\), the amplitude of the fundamental harmonic at the steady state, \(V_{o1}\), and the distortion factors \(HD_2\) or \(HD_3\). Hence, equating to zero the imaginary part of (7) (since \(V_{o1}\), \(a_1\) and \(a_3\) are real quantities) and finding at the next step the amplitude of the fundamental, gives respectively

\[
\omega_0 = \sqrt{\frac{1}{2LC + \left(\frac{L}{R}\right)^2}} \approx \sqrt{\frac{1}{2LC}}
\]

and

\[
V_{o1} = 2 \sqrt{\frac{1}{3a_3}} \left(\frac{2}{3\omega_0} - a_1\right) R
\]

where approximation in (8) holds if \(R^2 >> L(2C)\).

From second and third equations of (7), dividing by \(V_{o1}\), we find the harmonic distortion factors, \(HD_2\) and \(HD_3\) (the latter is given at the bottom of the page)

\[
HD_2 = \frac{X_2}{X_1} = \sqrt{2 |a_1|}\frac{L}{C} \times \\
\left[1 + \frac{2a_2}{a_1} \left(1 + \frac{3}{2} \frac{a_2}{a_1} V_{o2}^2 \right)^2 \right]^{-1}
\]

Equations (10) and (11) give useful design guidelines since they show how to control the amount of distortion due to the second and third harmonic.

\[
HD_3 = \frac{X_3}{X_1} = \frac{27}{8\sqrt{2}} |a_1| \sqrt{\frac{1}{Q}} \sqrt{\frac{9 + \frac{2}{Q} \left[1 + \frac{3}{2} \frac{a_2}{a_1} V_{o2}^2 - \frac{a_2^2}{a_1 a_3} V_{o3}^2 \right]^{-1}}{64 + 9 \left(1 + \frac{3}{2} \frac{a_2}{a_1} V_{o2}^2 \right)^2}} \times \\
\left[1 + 2a_2 \frac{1}{1 + \frac{3}{2} \frac{a_2}{a_1} V_{o3}^2} \frac{1}{a_3} \left(1 + \frac{3}{2} \frac{a_2}{a_1} V_{o2}^2 \right)^2 \right] V_{o3}^2
\]
3. Simulations

In order to verify the proposed model, we simulated a Hartley oscillator biased with $V_{DD}=3.0$ V and $V_{GG}=1.5$ V. For the MOSFET M1 we adopted the BSIM3V3 model provided by 0.35-µm AMS technology. Its aspect ratio was set equal to 10µm/1µm. Our target was a sinusoidal voltage whose amplitude and frequency were 500.0 mV and 100.0 MHz, respectively. In order to get the MOSFET nonlinear coefficients, we performed a Fourier analysis on the equivalent DC circuit whose schematic is shown in Fig. 1, replacing the two inductors by a short circuit and the capacitor by an open circuit, respectively. Indeed, these coefficients depend only on the quiescent point and on the amplitude of the fundamental, but not on the particular oscillation frequency. Hence we applied a sinusoidal voltage of the desired amplitude and frequency across the gate and source and measured the Fourier components of the drain current. In particular, from simulation they result equal to $a_1=846.1 \mu A/V$, $a_2=-307.1 \mu A/V^2$ and $a_3=67.8 \mu V^3$. From (9) and (8) it has to be $R=2.3$ kΩ and $LC=1.3 \times 10^{-8}$ s², respectively. It is worth noting that we have a degree of freedom in sizing the reactive components, since only their product affects the constraints. Hence, we can choose their values on the basis of properties such as the area occupied by the passive components or the THD referring to (10) and (11). In this example we simply set $L_1=L_2=10$ nH and $C=130$ pF.

We performed simulations with Spectre on the circuit shown in Fig. 1. Main simulation results are summarized in Table I, where we compared the amplitudes of harmonics ($V_{o1}, V_{o2}, V_{o3}$) and oscillation frequency ($f_0$) expected from our model and from simulation. From (10) and (11) we are able to evaluate $HD_2$ and $HD_3$ for the circuit simulated. Table II summarizes the values expected from our model and from simulation. The error is about 1 dB for both $HD_2$ and $HD_3$.

### Table I. Comparison between expected and simulated values of the output signal at the steady state, its harmonics, and the oscillation frequency for Hartley oscillator ($R=2.3$ kΩ, $L_1=L_2=10$ nH, $C=130$ pF)

<table>
<thead>
<tr>
<th></th>
<th>$V_{o1}$ (mV)</th>
<th>$V_{o2}$ (mV)</th>
<th>$V_{o3}$ (µV)</th>
<th>$f_0$ (MHz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>500.0</td>
<td>3.2</td>
<td>223.3</td>
<td>100.0</td>
</tr>
<tr>
<td>Simulated</td>
<td>476.0</td>
<td>3.4</td>
<td>235.3</td>
<td>99.8</td>
</tr>
<tr>
<td>Percentage error</td>
<td>4.8%</td>
<td>6.3%</td>
<td>5.4%</td>
<td>0.2%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$HD_2$ (dB)</th>
<th>$HD_3$ (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>-43.9</td>
<td>-67.0</td>
</tr>
<tr>
<td>Simulated</td>
<td>-43.0</td>
<td>-66.1</td>
</tr>
</tbody>
</table>

4. Conclusions

In this paper, we analytically studied a Hartley oscillator by using a novel approach that exploits the phasor’s method, which is a simple algebraic approach. The main linearity performance of the oscillator, namely the oscillation frequency, amplitude, and second- and third-order harmonic distortion factors, was established by a set of closed-form (algebraic) equations. The approach is particularly amenable from a design point of view, since it gives the designer a deeply insight into circuit behavior, unlike other methods like Volterra series and Poincaré’s perturbation which are more complex from a computational point of view. A Hartley oscillator was designed using the proposed technique, and simulated at the transistor level with Spectre. Simulations and expected results were found in very good agreement. Further application of the approach to other types of oscillators is currently in progress.

### Table II. Comparison between expected and simulated values of $HD_2$ and $HD_3$ for Hartley oscillator ($R=2.3$ kΩ, $L_1=L_2=10$ nH, $C=130$ pF)

### References


Amplitude response curves of frequency-locked rotations

Yoshihiko Susuki† and Yoshisuke Ueda‡
† Department of Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615–8510, Japan
‡ Department of Complex Systems, Future University–Hakodate
116–2 Kamedanakano, Hakodate, Hokkaido 041–8655, Japan
Email: susuki@dove.kuee.kyoto-u.ac.jp, ueda@fun.ac.jp

Abstract—This paper studies frequency-locked rotations in a phase-locked loop (PLL) circuit as FM demodulator. A rotation represents a desynchronized steady state in the PLL circuit and is regarded as another type of self-excited oscillations with natural rotation frequency. The rotation frequency can be locked at driving frequencies of modulation signals. This paper investigates amplitude response curves of frequency-locked rotations in the PLL circuit. The obtained results show that they have some different features from the well-known response curves of van der Pol oscillator.

1. Introduction

Phase-locked loops (PLLs) are well-known as an important technique in communication engineering [1]. The PLLs consist of phase comparators, loop filters, and voltage-controlled oscillators. Many researchers have studied nonlinear oscillations in PLL circuits [2, 3, 4]. The understanding of the oscillations is important for applying the PLL circuits to various engineering systems.

A system that is derived in [2, 3, 4] to represent the dynamics of a PLL circuit has two steady states. One is a stable equilibrium point which corresponds to the phase-locked steady state in the circuit. The other is a stable limit cycle resulting from periodic nature of phase space, called stable limit cycle of the second kind [5] or stable rotation [6]. The rotation represents a desynchronized steady state in the PLL circuit and is regarded as another type of self-excited oscillations with natural rotation frequency. Such rotating steady state is studied in the PLL circuit [2] to analyze its pull-in range and is also examined in the Josephson junction [6] to clarify its current-voltage characteristics.

The present paper considers frequency-locked rotations in the PLL circuit as a FM demodulator. Frequency entrainment or locking by external force is well-known in periodically driven self-oscillatory systems [7, 5, 8, 9]. In the van der Pol oscillator [7], we focus on stable limit cycles of the first kind [5] or stable librations [6] and examine locked librations at driving frequencies of external force to the oscillator. Amplitude response curves in [7, 5, 8] show a characteristic feature of the frequency-locked librations. On the other hand, in the case of the PLL circuit with stable rotation, the rotation frequency can be also locked at driving frequencies of modulation signals. The similar phenomenon is actually discussed in Josephson Junction circuits [10, 11, 12]. It is important to clarify the frequency-locked rotations for the evaluation of pull-in range of the PLL circuit under modulation signals. This paper shows amplitude response curves of frequency-locked rotations in the PLL circuit. The obtained results show that they have some different features from the well-known response curves of van der Pol oscillator.

2. Mathematical model

This section introduces a mathematical model which describes the dynamics of the PLL circuit as a FM demodulator [3]. The PLL circuit includes a sinusoidal phase comparator, a lead-lag loop filter, and a voltage-controlled oscillator. Suppose that \( \phi \) denotes the phase error of input signals to the phase comparator, then the phase error dynamics are represented by the following system of a second-order differential equation:

\[
\begin{align*}
\frac{d\phi}{dt} &= y, \\
\frac{dy}{dt} &= -\beta y + \beta \sigma - \sin \phi \\
&\quad + m\sqrt{\beta^2 + \Omega^2} \cos \Omega t.
\end{align*}
\]

(1)

\( \beta \) and \( \sigma \) are the fixed parameters, and \( m \) and \( \Omega \) are the parameters of modulation signal. The parameters are in per unit system. The derivation of the system (1) is given in [3]. The driving term \( m\sqrt{\beta^2 + \Omega^2} \cos \Omega t \) is slightly modified from the original one [3] for simplicity of the present analysis. The two variables \( (\phi, y) \) in the system (1) belong to cylindrical phase plane because of the periodic restoring term. The periodic nature of phase space results in the occurrence of rotating steady states. Additionally, to derive theoretical response curves of frequency-locked rotations in Section 3, define a smooth function \( S(\phi, y) \) for the system.
Figure 1: Waveform of stable rotation in the system (1) under \( m = 0 \)

\[
(1) \quad S(\phi, y) = \frac{1}{2} y^2 - \cos \phi - \beta \sigma \phi. \tag{2}
\]

The following equality is then obtained for any solution \((\phi(t), y(t))\):

\[
S(\phi(t), y(t)) - S(\phi(0), y(0))
\]

\[= -\beta \int_0^t \{y(\tau)\}^2 d\tau \]

\[+ \int_0^t y(\tau) m \sqrt{\beta^2 + \Omega^2} \cos \Omega \tau d\tau. \tag{3}
\]

3. Amplitude response curves

This paper shows amplitude response curves of frequency-locked rotations under the parameters \( \Omega \) and \( m \). The system (1) under \( m = 0 \) and \(|\beta \sigma| > 4\beta / \pi\) has one stable limit cycle of the second kind with rotation angular frequency \( \Omega_0 \). The condition of \( \beta \) and \( \sigma \) is obtained with the Melnikov’s method [13]. The corresponding solution \( \phi_0(t) \) is represented using the periodic function \( x_0(t) \) of \( 2\pi/\Omega_0 \) as follows:

\[
\phi_0(t) = \Omega_0 t + x_0(t). \tag{4}
\]

Fig. 1 shows the waveform of stable rotation in the system (1) under \( m = 0 \) and the following parameters settings:

\[
\beta = 0.56, \quad \sigma = 1.7. \tag{5}
\]

The setting is based on [4]. \( \Omega_0 \) is then 1.6. Within the small difference between natural rotation frequency \( \Omega_0 \) and driving frequency \( \Omega \), it is expected that we observe a locked rotation at the driving frequency \( \Omega \). The type of frequency entrainment is called harmonic entrainment.

The averaging method [14] is here used for the derivation of response curves. For the harmonic entrainment, the corresponding solution \( \phi(t) \) is expressed via the periodic function \( x(t) \) of \( 2\pi/\Omega \) as follows:

\[
\phi(t) = \Omega t + x(t). \tag{6}
\]

It is now assumed that the solution \( x(t) \) is approximately expressed as follows:

\[
x(t) = \frac{A_0}{2} + A_1 \cos(\Omega t + \varphi_1) \tag{7}
\]

where \( A_0 \) is constant, and \( A_1 \) and \( \varphi_1 \) are functions of \( t \) which are only slowly variable. By substituting the above solution \( x(t) \) and \( y(t) = d\phi(t)/dt = \Omega - \Omega A_1 \sin(\Omega t + \varphi_1) \) into the system (1) and using the first-order averaging method [14], we have the following averaged system:

\[
\begin{align*}
\frac{dA_1}{dt} &= \frac{1}{2\Omega} \left[ (J_0(A_1) + J_2(A_1)) \cos \left( \frac{A_0}{2} - \varphi_1 \right) \right. \\
&\quad - \beta \sqrt{\beta^2 + \Omega^2} \sin \varphi_1 - \beta \Omega A_1 \right], \\
A_1 \frac{d\varphi_1}{dt} &= \frac{1}{2\Omega} \left[ (J_0(A_1) - J_2(A_1)) \sin \left( \frac{A_0}{2} - \varphi_1 \right) \right. \\
&\quad - \beta \sqrt{\beta^2 + \Omega^2} \cos \varphi_1 - \Omega^2 A_1 \right],
\end{align*}
\]

where \( J_n \) for \( n = 0, 2 \) are the Bessel functions of the first kind.

Theoretical response curves of frequency-locked rotations are derived through the above preliminary. The three variables \((A_0, A_1, \varphi_1)\) are required for the derivation. Namely we need to obtain three independent equations. Two of the required equations are induced by equating the right-hand sides of the averaged system (8) to zero. The remaining equation is obtained from eq. (3) of the original system (1). For the harmonic entrainment, \( \phi(t) \) and \( y(t) \) obviously satisfy the following periodic property:

\[
\phi \left( \frac{2\pi}{\Omega} \right) = \phi(0) + 2\pi, \quad y \left( \frac{2\pi}{\Omega} \right) = y(0). \tag{9}
\]

By substituting the above property, the solution (7), and the corresponding variable \( y(t) \) into eq. (3), we have the following equation for the derivation of response curves:

\[
-2\pi \beta \sigma = -\pi \beta \Omega (2 + A_1^2) \]

\[-\pi m \sqrt{\beta^2 + \Omega^2} A_1 \sin \varphi_1. \tag{10}
\]

Figure 2 shows several response curves for harmonic amplitude of frequency-locked rotations under the parameters (5). The figure (a) shows the harmonic amplitude \( A_1 \) and is obtained by plotting the above three equations in \( \Omega - A_1 \) plane for several values of \( m \).
Fig. 2 (b) describes the harmonic amplitude of $x(t)$ and is given with numerical integration of the original system (1). The response curve at $m = 0.6$ in the figure (a) is described in the absolute value. Each closed curve in the figure (a) contains amplitude response of both stable and unstable periodic solutions. On the other hand, in the figure (b) with numerical integration, the solid lines describe the response curves for stable solutions, and the broken ones for saddle-type solutions. The symbol $\times$ in the figure (b) denotes the parameters set $(\Omega, m) = (\Omega_0, 0)$. The corresponding value is equal to the harmonic amplitude of $x_0(t)$ for the stable rotation in Fig. 1. Fig. 2 (b) here suggests that the harmonic amplitude of $x(t)$ is precisely zero at the parameters set $(\Omega, m) = (1.7, 0.5587)$. The corresponding waveform of $x(t)$ is shown in Fig. 3(b) and suggests that all the harmonics of $x(t)$ have zero amplitude. Fig. 2 implies that the theoretical result with averaging method almost corresponds to that with numerical integration.

**4. Discussion**

This paper showed amplitude response curves of frequency-locked rotations in the PLL circuit. Last of all, the obtained response curves are considered by comparison with frequency-locked librations. The response curves in van der Pol oscillator are here adopted for the comparison, which are given in Fig. 2 of [7], Fig. 113.1 of [5], or Fig. 12.1 of [8]. The shape of response curves in Fig. 2 differs from that in van der Pol oscillator. In the van der Pol oscillator, stable solutions have the maximum amplitude near the zero detuning, that is, the natural frequency. On the other hand, in the PLL circuit, every harmonic amplitude for stable solutions attains the minimum value near the natural rotation frequency $\Omega_0$. Furthermore it gradually increases to each side of the minimum value in a different manner from the response curves in van der Pol oscillator.
Pol oscillator. This is supported by the waveforms of $x(t)$ in Fig. 3. The obtained response curves in this paper hence delineate some different features between the frequency-locked rotations and librations.

**Acknowledgment**

The first author greatly appreciates fruitful discussion with Professor Takashi Hikihara at Kyoto University.

**References**


Statistical physics in information theory: an overview

Ljupco Kocarev† and José M. Amigó‡

†Institute for Nonlinear Science, University of California San Diego, La Jolla, CA, USA, and Macedonian Academy of Sciences and Arts, Skopje, Macedonia
‡Centro de Investigación Operativa, Universidad Miguel Hernández, 03202 Elche, Spain
Email: lkocarev@ucsd.edu, jm.amigo@umh.es

Abstract—An overview of analytic tools for design, analysis, and modeling communication systems described by linear vector channels, when the number of components in each vector is large, is presented. The replica method developed in statistical physics for analyzing spin glasses is reviewed from the viewpoint of its applications in communications engineering.

1. Introduction

In recent years, statistical physics has been providing important concepts and powerful analytical tools in studying various problems related to probability-based information processing [1], such as: Hopfield neural network model of associative memory [2], statistical learning of perceptrons [3], low-density parity-check (LDPC) codes [4], and code-division multiple-access (CDMA) multiuser detectors [5]. In a communication system information is transmitted from various sources to various sinks via a common medium called a channel. Examples of such systems include telephone networks, both fixed and wireless, the Internet, local area networks, and so on. The complexity of communication systems increases with the number of people and/or data streams to be handled simultaneously. This complexity affects the design, the modeling, and the analysis of the system. From an engineering point of view, it is particularly important to be able to predict the behavior of the system, before it is actually built.

Physicists have successfully developed statistical physics to explain the evolution of macroscopic quantities such as temperature and pressure, though the microscopic behavior of the molecules is only described statistically. Communication systems for multiple data streams can be modeled as well by statistical interactions between the signals belonging to different data streams. Provided that the number of data streams transmitted simultaneously through the system is large enough, similar effects as in statistical physics occur. In this paper we overview some statistical physics’ tools for design, analysis, and modeling communication systems, when the number of components is large.

2. Random Matrix Theory

Consider linear vector memoryless channels of the form

$$\mathbf{y} = \mathbf{H} \mathbf{x} + \mathbf{n},$$

(1)

where \(\mathbf{x}\) is the \(K\)-dimensional input vector, \(\mathbf{y}\) is the \(N\)-dimensional output vector, the \(N\)-dimensional vector \(\mathbf{n}\) models the additive circularly symmetric Gaussian noise, and \(\mathbf{H}\) is the \(N \times K\) random channel matrix. All these quantities are, in general, complex-valued. The nature of the \(\mathbf{x}, \mathbf{y}\) and \(\mathbf{H}\) depends on the actual application. For example, in CDMA, the components of the vector \(\mathbf{x}\) are regarded as the signals of \(K\) individual users while the matrix \(\mathbf{H}\) contains their spreading sequences as columns. In antenna array communications, the components of the vectors \(\mathbf{x}\) and \(\mathbf{y}\) represent the signals sent and received by the \(K\) transmit and \(N\) receive antenna elements, respectively. In orthogonal frequency-division multiple access (OFDM), the components of the vectors \(\mathbf{x}\) represents the \(K\) subcarriers at transmitter, and the matrix \(\mathbf{H}^\mathsf{T} \mathbf{H}\) accounts for inter-carrier interference.

Random matrix theory has found many applications in physics, statistics and engineering, including stochastic differential equations, condensed matter physics, statistical physics, chaotic systems, numerical linear algebra, neural networks, multivariate statistics, information theory, and signal processing. The first asymptotic results on the limiting spectrum of large random matrices were obtained by Wigner in the 1950s in a series of papers motivated by nuclear physics [6]. Wigner initially dealt with an \(n \times n\) symmetric matrix \(\mathbf{A}\) whose diagonal entries are 0 and whose upper-triangle entries are independent and take the values \pm 1 with equal probability. Wigner showed that, as \(n \to \infty\), the averaged empirical distribution of the eigenvalues of \((1/\sqrt{n})\mathbf{A}\) converges to the semicircle law whose density is

$$w(x) = \begin{cases} \frac{1}{\pi \sqrt{4-x^2}} & \text{if } |x| \leq 2 \\ 0 & \text{if } |x| > 2. \end{cases}$$

If all entries of \(\mathbf{A}\) are chosen to be i.i.d., then the eigenvalues of \((1/\sqrt{n})\mathbf{A}\) are asymptotically uniformly distributed on the unit circle of the complex plane. This
is commonly referred to as Girko’s full-circle law [6]. The asymptotic theory of singular values of rectangular matrices has concentrated on the case where the matrix aspect ratio converges to a constant $K/N \rightarrow \beta$, as the size of the matrix grows. The first success in the quest for the limiting empirical singular value distribution of rectangular random matrices is due to Marčenko and Pastur in 1967 [6]. A central result in random matrix theory states that when the entries of $H$ are zero-mean i.i.d. with variance $1/N$, the empirical distribution of the eigenvalues of $HH^\dagger$ converges almost surely, as $K, N \rightarrow \infty$, with $K/N \rightarrow \beta$, to a non-random limit whose density function is

$$f_\beta(x) = (1 - \beta)^+ \delta(x) + \frac{\sqrt{(x-a)^+(b-x)^+}}{2\pi x}$$

where $(z)^+ = \text{max}(0, z)$, $a = (1 - \sqrt{\beta})^2$ and $b = (1 + \sqrt{\beta})^2$.

If we assume an i.i.d. Gaussian input, the normalized input-output mutual information of (1) conditioned on $H$ is [6]:

$$\frac{1}{N} I(x; y | H) = \frac{1}{N} \log \det(I + \text{snr}HH^\dagger)$$

$$= \frac{1}{N} \sum_{i=1}^{N} \log \left( 1 + \text{snr} \lambda_i(HH^\dagger) \right),$$

where $\text{snr}$ is the transmitted signal-to-noise ratio:

$$\text{snr} = \frac{NE(\|x\|^2)}{K^2E(\|n\|^2)}$$

and $\lambda_i(HH^\dagger)$ is the $i$th squared singular value of $H$. Equation (2) represents the capacity of the synchronous DS-CDMA channel as a function of the signature vectors (Verdú, 1986) [6]. Therefore, the capacity depends on the distribution of the empirical (squared) singular value distribution of the random channel matrix. A rich body of results exists analyzing the asymptotic spectrum of $H$ as the number of columns and rows goes to infinity while the aspect ratio of the matrix is kept constant.

In CDMA, channels with $K$ and $N$ between 32 and 64 would be fairly typical. Surprisingly, even quite smaller system sizes are large enough for the asymptotic limit to be an excellent approximation. Furthermore, not only do the average of (2) converges to its limit surprisingly fast, but the randomness in this functional due to the random outcome of $H$ disappears extremely quickly. The convergence of the singular values of $H$ exhibits several key features [6]: (i) Insensitivity of the asymptotic eigenvalue distribution to the shape of the p.d.f. of the random matrix entries. (ii) Ergodic behavior: the eigenvalue histogram of any matrix realization converges almost surely to the average asymptotic eigenvalue distribution. (iii) Fast convergence of the empirical singular-value distribution to its asymptotic limit. For other applications of random matrix theory to communication see, for example, [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17].

3. Spin glasses and replica method

In the previous section, we consider eigenvalues (and eigenvectors) of random matrices. However, in order to analyze and design large dimensional communication systems which depend on more complicated functions of the channel matrix, such as minimum distance between signal points, a more powerful machinery than random matrix theory is needed. Such a machinery was developed in statistical physics for the analysis of some particular magnetic materials called spin glasses and is known as the replica method [19].

The replica method was introduced by Mezard and Parisi [19]. They employed the replica method to study the $n$th moment of the partition function in the limit $n \rightarrow 0$. The basic idea is that physical quantities are derived from the disorder-averaged free energy

$$\mathcal{F} = -T \ln \mathbb{E}[Z]$$

It is difficult in practice to carry out such an average of a logarithm, but it may be accomplished by using an amusing mathematical identity:

$$\mathcal{F} = -T \lim_{n \rightarrow 0} \frac{Z^n - 1}{n}$$

Technically, one proceeds by noticing that $Z^n$ for integer $n$ can be written as a field theory involving $n$ copies or replicas of the original system, described by fields $u^\alpha$, with $\alpha = 1, \ldots, n$. Performing the disorder average then leads to an interacting system,

$$Z^n = \langle \int d(u^\alpha)e^{-S} \rangle$$

with

$$S = \int d^d u \left\{ \frac{\sigma}{2T} \sum_{\alpha} (\nabla u^\alpha)^2 - \frac{1}{2T^2} \sum_{\alpha, \beta} R(u^\alpha - u^\beta) \right\},$$

where we have taken a Gaussian random potential with a two-point correlation function of the form

$$V(u, x)V(u', x') = \delta^d(x - x')R(u - u').$$

Various forms may be taken for $R$, depending upon whether one prefers to study short-range or long-range correlated disorder.

We illustrate the replica method on LDPC codes. LDPC codes were originally introduced by Gallager in 1962 [4]. They rely on a sparse linear transformation of binary messages at the decoding stage, while encoding relies on a dense matrix generated by the inverse
of the sparse linear transformation. Gallagers code is defined by a binary matrix \( H = [A|B] \), concatenating two very sparse matrices known to both sender and receiver, with \( B \) (of dimensionality \((N-K)\times(N-K)\)) being invertible and \( A \) of dimensionality \((N-K)\times K\). The matrix \( H \) can be either random or structured, characterized by the number of non-zero elements per row/column.

Encoding refers to the mapping of a \( K \)-dimensional binary vector \( s \in \{0,1\}^K \) to \( N \)-dimensional codewords \( t \in \{0,1\}^N \) \((N > K)\) by the linear product \( t = G^T s \mod(2) \). To demonstrate the way in which Gallager’s code is utilized we consider the BSC, where the encoded vector \( t \) is corrupted by a noise vector \( n \in \{0,1\}^N \) with components independently drawn from \( P(n) = (1-p)\delta(n) + p\delta(n-1) \). The received vector takes the form \( r = G^T s + n \mod(2) \). Decoding is carried out by multiplying the received message by the matrix \( H \) to produce the syndrome vector \( z = Hr = Hn \mod(2) \). Decoding refers to finding an estimate of \( n \) knowing \( z \) and \( H \); this, of course, enables one to obtain the original message vector \( s \) (the first \( K \) bits of \( r + n \mod(2) \)). Maximum a posteriori (MAP) estimator is based on selecting the noise vector of the lowest weight that obeys all parity checks; this corresponds to mapping the received vector onto the nearest codeword. The noise vector MAP estimator, which is also the maximum likelihood (ML) estimator of the codeword, minimizes the block error probability, but is computationally demanding and cannot be used in practice. In practice, decoding is carried out mainly by employing some message passing algorithm such as belief propagation (BP) and its variations [4].

The link between error correcting codes and statistical mechanics was first pointed out by Sourlas [4]. The similarity between Ising spin models and LDPC codes stems from the formulation of the decoding problem. Employing the isomorphism between the additive Boolean group \((\{0,1\},\oplus)\) and the multiplicative binary group \((\{+1,-1\},\odot)\), whereby every addition in the Boolean group corresponds to a unique product in the binary group and vice versa, one can map the decoding problem to a Gibbs distribution by constructing an appropriate Hamiltonian. The decoding problem depends on posteriors such as \( P(\tau|r) \), where \( r \) is the observation (received message or syndrome vector), and \( \tau \) is a candidate estimate of the unknown original message \( s \) (or alternatively a candidate noise vector from which an estimate of the noise can be obtained). Applying Bayes theorem this posterior takes the form \( P_{\sigma|\gamma}(\tau|r) = \frac{1}{Z(r)}\exp[\ln P_{\sigma}(r|\tau) + \ln P_{\sigma}(\tau)] \) (15) where \( \sigma \) and \( \gamma \) are hyper-parameters assumed to describe features such as the encoding scheme, source distribution and noise level. This form suggests the following family of Gibbs measures (\( \beta \) being the inverse temperature):

\[
P_{\alpha,\beta}(\tau|r) = \frac{1}{Z} \exp[-\beta H(\tau;r)]
\]

\[
H(\tau;r) = -\ln P_{\sigma}(r|\tau) - \ln P_{\sigma}(\tau)
\]

The received corrupted codeword depends on the coding mechanism and channel noise, both of which represent the quenched disorder in the system. The MAP estimator of \( s \) is clearly obtained at the ground state of the Hamiltonian, i.e. by the sign of thermal averages \( s^{MAP}_j = \text{sign}(\langle \tau_j \rangle_{\beta \to \infty}) \) at zero temperature.

A key point is the definition of an appropriate Hamiltonian. In the case of a BSC, the Hamiltonian takes the form [4]:

\[
H_\gamma(\tau;n) = -\gamma \sum_{i_1\cdots i_k} D_{i_1\cdots i_k}(\tau_{i_1} \cdots \tau_{i_k} - 1) - F \sum_{i=1}^N n_i \tau_i,
\]

where \( F = (1/2)\ln(1-p)/p \) and the tensor \( D \) represents the connectivities of the matrix \( H \). Once the Hamiltonian has been defined one can calculate the free energy of the system and study emerging solutions for various choices of the parameters and levels of channel noise. Two main methods can be employed for carrying out the analysis, the replica method for diluted systems and the Bethe approximation [4]. Analysing the typical performance of Gallager codes is based on similar studies of diluted systems. The aim is to compute the free energy:

\[
F = -\frac{1}{\beta} \lim_{N \to \infty} \frac{1}{N} \langle \ln Z/H \rangle_{\tau} n,
\]

where \( Z = \text{Tr}_r \exp(-\beta H(\tau;n)) \), from which the typical macroscopic (thermodynamic) behavior can be obtained using the Hamiltonian \( H \). Quenched averages are carried out over the connectivity tensor \( D \) and the true noise vector \( n \), see [4] for details. Phase transitions occur in spin glass systems. If a system parameter changes, the free energy may shift its favor from the present to another solution. Since each solution corresponds to a different macroscopic behavior of the system, changing the valid solution means that a phase transition takes place. Phase transitions have also been found in turbo decoding and detection of CDMA, see [18] and [5], respectively. For other applications of replica methods to communication see [20, 21, 22, 23, 24, 25, 26].

**References**


653


Statistical physics of information flow in wireless networks

Massimo Franceschetti†
†Dept. of Electrical and Computer Engineering
University of California at San Diego
La Jolla, CA 92093-0407
massimo@ece.ucsd.edu

Abstract—We give an overview of the statistical physics approach to model the information flow in wireless networks. By exploiting properties of random graphs, we characterize the dynamics of the network information flow.

1. Introduction

Spatial stochastic models of networks [1, 2] have recently generated great interest, due to their ability to describe wireless ad-hoc and sensory/communication systems. Nodes in such systems are often deployed at random in a given area and self assemble into a wireless communication network structure. Therefore, it is widely recognized that random geometric graphs theory can provide the fundamental analysis and design tools required in this context.

However, often random geometric graph and percolation-like models suffer from an inherent static description of the graph structure. In considering communication over a spatial structure dynamic issues of interference and cooperation between randomly located nodes arise and must be dealt with. The notion of communication link itself must be redefined, since in wireless there are no static links a priori, as nodes can share the medium in arbitrarily complex ways. Based on these observations, after a brief overview of relevant statistical physics tools, we describe a new approach [7, 8, 9] that extends classic random network models to provide a dynamic description of information transfer over random structures.

2. Percolation theory

In its original formulation [3], continuum percolation theory considered wireless transmitting stations of range $2r$ randomly located on the plane, according to a two-dimensional Poisson point process, and formulated the problem of whether the system can provide some long distance multi-hop communication. Using a branching process argument, Gilbert [3] showed the existence of a critical value $\lambda_c(r) = \lambda_s(1)/r^2$ for the density of the transmitters, such that, for $\lambda > \lambda_c$, an unbounded connected component of transmitters forms (i.e., the network percolates), with probability one, and so long-distance multi-hop communication is possible. On the contrary, for $\lambda < \lambda_c$ any connected component is bounded, see Fig. 1. The exact critical value of $\lambda_c$ is still unknown [1], but it can be estimated by numerical simulation to be $\lambda_c \approx 0.359$ for unit discs. This corresponds to an expected number of connec-

Figure 1: Phase transition. In this simulation discs of radius $r = 1$ are centered at each random point. These represent broadcasting stations, so that if two discs overlap the corresponding stations are within range and they are part of the same connected component. Connected components are identified by the same color. On the left-hand side of the figure, the density of the discs is $\lambda = 0.3$; on the right-hand side $\lambda = 0.4$. The unbounded connected component forms at the critical value $\lambda_c \approx 0.359$, corresponding to an average node degree of 4.51.
tions per node

\[ \text{ENC} = \lambda_c(r)4\pi r^2 = \frac{\lambda_c(1)4\pi r^2}{r^2} \approx 4.51, \]

independent of the connection radius \( r \).

The main idea behind Gilbert’s theory is to consider the circular geometries of the discs in Figure 1 as radiation patterns of signals transmitted by the Poisson points. Consider two points of the Poisson process and label them a transmitter \( x \) and a receiver \( y \). The transmitter \( x \) radiates a signal with intensity proportional to the power \( P \) spent to generate the transmission. The signal diffuses isotropically in the environment and is then received by \( y \) with intensity \( P \) times a loss factor \( \ell(x,y) \leq 1 \), due to isotropic dispersion and absorption in the environment. Furthermore, the reception mechanism is affected by noise, which means that \( y \) is able to detect the signal only if its intensity is sufficiently high compared to the environment noise \( N > 0 \). We conclude that \( x \) and \( y \) are able to establish a communication link if the signal to noise ratio (SNR) at the receiver is above a given threshold \( \beta \). That is, if

\[ \text{SNR} = \frac{P(\ell(x,y))}{N} > \beta. \]  

It is reasonable to assume the loss factor to be a decreasing function of the Euclidean distance between \( x \) and \( y \). It follows that fixing the threshold \( \beta \) is equivalent to fixing the radius \( r \) of the discs in Gilbert’s model.

An immediate application of percolation theory is in determining the critical number of connections per node needed in a wireless network to propagate information from one side to the other of a massively large network. Moreover, the ENC is also an indicator of robustness, as it shows how many node failures the network can tolerate, before breaking down into a myriad of disconnected clusters. Furthermore, the theory has found applications in the physics of materials, biology, and epidemiology [1].

3. Random graphs with interference

From a practical viewpoint, however, percolation theory based models suffer from the inherent geometric disc abstraction; namely transmission is assumed to be isotropic, and reliable communication channels are assumed to exist between nodes closer than a given distance. A first step towards a more practical generalization of this model has been provided in [4], and later extended in [5], where a dependent percolation model has been introduced, in which interference between nodes is explicitly taken into account when considering their connectivity properties. In previous percolation models of networks links represent the possibility of direct transmission between pairs of nodes at a given time, but the model does not account for the possible interference due to simultaneous transmission of other nodes in the network. In this case, all nodes can contribute to the amount of noise present at the receiver. Consider two points of a planar Poisson point process \( x_i \) and \( x_j \), and assume \( x_i \) wants to communicate with \( x_j \). At the same time, however, all other nodes \( x_k \), \( k \neq i \), also transmit an interfering signal that reaches \( x_j \). We can write the total interference term at \( x_j \) as \( \gamma \sum_{k \neq i} P(\ell(x_k, x_j)) \), where \( P \) is the transmitted power, and \( \gamma > 0 \) is a factor that depends on the technology adopted in the system. Accordingly, in [4, 5] the simplistic disc model has been modified into a signal to noise plus interference ratio (SNIR) model by which \( x_i \) and \( x_j \) can establish a communication link if

\[ \text{SNIR} = \frac{P(\ell(x_i, x_j))}{N_0 + \gamma \sum_{k \neq i} P(\ell(x_k, x_j))} > \beta. \]

A random network is then obtained by computing the \( \text{SNIR} \) at each pair of Poisson points and drawing an edge if this exceeds the threshold \( \beta \). This network clearly does not have the full independence structure of Gilbert’s model, because the presence of an edge between any pair of nodes now depends on the random positions of all other nodes in the plane that are causing interference, and not only on the two end-nodes of the link. Such dependencies make the mathematical analysis of these kind of networks considerably more challenging.

4. Information flow

One more key step is required to push the work of [4, 5] further, and to move from connectivity models to communication flow models. We note that the models described up to now considered only the existence and Euclidean length of the links in the network. The concept of communication was simply tied to the concept of connectivity: communication between pairs of nodes is successful if the receiver can detect a transmitted signal, despite the attenuation due to the distance between the nodes, the noise due to the physical act of transmission, and the interference due to other simultaneous transmissions. A more general view arises if one considers the information rate at which communication can be performed in the random network, rather than connectivity alone. In fact, by information theory arguments it is well known that it is theoretically possible for any SNIR between two nodes, to have a non-zero communication rate. Hence, any pair of nodes is always connected, but its rate is limited by the random geometric configuration of the nodes.

To be more precise, assume that all the nodes in the network are simultaneously transmitters and receivers, and each transmitter wishes to communicate to some receiver. What is an achievable rate between a pair of nodes \( x_i \) and \( x_j \) in this case? It turns out that this rate depends on the random spatial configuration of the nodes in the network, that can be analyzed with the original percolation theory model. Recall from our discussion above, that a signal of intensity \( P \) radiated isotropically from \( x_i \) to \( x_j \) is, in first approximation, subject to a loss factor \( \ell(x_i, x_j) \), due to diffusion and absorption in the environment. Then, by treating
simultaneous transmission of other nodes in the network causing interference on the same footing as random noise, an achievable rate between $x_i$ and $x_j$ can be computed by using Shannon’s formula,

$$R = \log \left( 1 + \frac{P_l(x_i, x_j)}{N_0 + \sum_{kl} P_l(x_k, x_l)} \right),$$

By (3) we see that every pair of nodes has a rate that depends not only on their relative position, but also on the (random) positions of all other nodes in the network. Moreover, different operating strategies lead to different rates. For example, all nodes beside $x_i$ and $x_j$ could be kept silent, while $x_i$ transmits to $x_j$, so that (3) reduces simply to $\log(1 + SNR)$; or some nodes could be used to relay information in a multi-hop fashion from $x_i$ to $x_j$, in which case the end-to-end rate between $x_i$ and $x_j$ corresponds to the smallest achievable rate along the multi-hop chain. Therefore, we conclude that the geometry of a random information flow network model is that of an infinite clique, where vertices are Poisson points and the achievable flow on each link depends by the random spatial configuration of the nodes and by their transmission strategies.

Now, by taking the statistical physics approach, a natural question to ask is how the achievable information rate per source destination pair in the random network scales with the system’s size. To compute such throughput scaling laws that are valid for all source destination pairs, we need to describe a specific transmission strategy that (i) selects some links in the clique and (ii) using these links, it provides end-to-end flow to all source destination pairs at the desired rate. Percolation theory can be used to determine the existence of such links having the desired properties, with high probability as the system size scales. Hence, the novelty of the approach is in exploiting the static spatial structure of the nodes distribution, given by percolation theory, to compute network dynamics in terms of information flows.

We point out that throughput scaling laws were first introduced by Gupta and Kumar in [6] and quickly sparked a wide range of attention in the research community [10, 11, 12, 13, 14, 15, 16]. Gupta and Kumar have proven that, under a certain network model, the per-node transmission rate decreases to zero as $\sqrt{1/n} \log n$, as the system size increases and more and more users need to share the wireless medium. Following up on this work [9], we were able to obtain an improved bound of $\sqrt{1/n}$ on the achievable information rate using our percolation approach. Furthermore, in the case when only two nodes in the network wish to communicate, and all others act as relays (wireless relay networks), we have also shown [7] how the information flow is related to the alpha-connectivity property of the percolation cluster.

It is also important to note that achievable scaling laws that can be shown constructively, using percolation theory and given routing strategies, are limited by corresponding upper bounds. Typically such upper bounds are obtained using an information theoretic cut-set type bound such as in [16]. Although we obviously cannot surpass such upper bounds on achievable rates using the percolation approach, we can still try to design strategies that will be able to improve the multi-hop delay and/or energy consumption for some communication patterns. Moreover, the question of exploiting the global percolation structure of the network using only local knowledge at each node is still open.

5. Conclusion

Percolation theory is known to be a powerful tool to model structural properties of networks. It has been shown that by exploiting relevant properties of the spatial structure of the nodes distribution it is also possible to compute network dynamics in terms of information flows. The reader is referred to the works [7, 9] where more technical details are spelled out.

References


Network structural properties, communication models and traffic dynamics

Mario di Bernardo∗◦, Francesco Sorrentino‡◦

Abstract—In this paper, we expound some of our recent results concerning the characterization of the relationship between the network topology and traffic dynamics taking place over it [1], [2]. We use a model of network generation that allows the transition from random to scale free networks. Specifically, we consider three different topological types of network: random, scale-free with $\gamma = 3$, scale-free with $\gamma = 2$. We firstly compare the performance of these networks in terms of throughput and average delivery time, under the hypothesis of constant transmission rates and infinite queue lengths at the network vertices. Interestingly, scale-free networks that are characterized by shorter characteristic-path-length (the lower the exponent, the lower the path-length), show worst performances in terms of communication. We present an explanation of this in terms of changes in the load distribution, defined as the number of shortest paths going through a given vertex. Then the issue is addressed of how the traffic behavior on the network is influenced by the variable factors of the transmission rates and queue length restrictions at the network vertices. We show that these factors can induce drastic changes in the throughput and delivery time of the network and are able to counter-balance some undesirable effects due to the topology.

I. Introduction

Much research effort has been spent recently in understanding the relationship between network topological features and communication performances. As a first approximation, it would be natural to make the most general hypothesis about the structure of the underlying network, that is, to think of it as a random graph. Unfortunately, real networks show statistical properties that are far from being completely random. The most important difference is that they have typically power law degree distributions with exponents between 2 and 3 [3]. Thus, in what follows, we consider three different topologies, in the order: random, scale-free with $\gamma = 3$, scale-free with $\gamma = 2$.

By making use of a packet transport model that has been widely studied in the literature (see [4], [5], [6] for further details), we compare the main indicators of the network performance, specifically the delivery time and the number of delivered packets (or throughput), as the underlying topology is varied.

II. Network generation model

In order to cause the transition from random to scale-free network we use the static model recently introduced in [7]. Vertices are indexed by an integer $i$, for ($i = 1, ..., N$), and assigned a weight or fitness $p_i = i^{-\alpha}$ where $\alpha$ is a parameter between 0 and 1. Two different vertices are selected with probabilities equal to the normalized weights, $p_i/\sum_k p_k$ and $p_j/\sum_k p_k$ respectively and an edge is added between them unless one exists already. This process is repeated until $M$ edges are made in the system leading to the mean degree $\langle k \rangle = 2M/N$. This results in the expected degree at vertex $i$ scaling as $k_i \sim (\frac{N}{M})^{\alpha}$ [7]. We then have the degree distribution, i.e. the probability of a vertex being of degree $k$, given by $P(k) \sim k^{-\gamma}$ with $\gamma = 1 + \frac{1}{\alpha}$. Thus, by varying $\alpha$, we can obtain the exponent $\gamma$ in the range, $2 < \gamma < \infty$. Moreover the ER graph is generated by taking $\alpha = 0$.

It is worth noting that the static model described here, can be considered as an extension of the standard ER model for generating random-scale free networks, i.e. networks with prescribed degree distribution, but completely random with respect to all the other features.

III. Load distribution in Networks

One of the main parameters of vertex centrality is betweenness centrality defined as the number of shortest paths between pairs of nodes crossing a given vertex [8]. Taking this index as a starting point, Goh et al. [7] [9], defined the load at each vertex $v$, say $l(v)$, as the number of packets passing through it, under the assumption that every node sends a packet to every other node in the network and that packets move in parallel from origin to destination through the geodesic, i.e. the shortest path between them. This implies that for each shortest path between a given couple of vertices, there is a packet passing along it; in the case that packets encounter a branching point at which there is more than one shortest path toward the destination, they would be divided by the number of branches at the branching point.

As pointed out in [1], in comparison to random graphs, scale free networks are characterized by:

- Lower average load (averaged over all the network vertices).
- Higher load standard deviation.

Intuitively, the presence of hubs in scale free networks, results in a shorter average distance between vertices. On the other hand, the increase in the load standard deviation indicates that this happens at the expense of the fairness of the network resources exploitation, with a relatively few vertices drawing most of the network traffic. As we will show next, such a phenomenon results to be particularly noxious to the communication dynamics taking place over the network.
IV. Model of Network Data Traffic

We use the family of Erramilli interval maps as the generator for each LRD traffic source, (Erramilli et al., 1994)[10] within the network, as further explained in [1].

We assume, the network involves two types of nodes: hosts and routers. The first are nodes that can generate and receive messages and the second can only store and forward messages. The density of hosts $\rho \in [0,1]$ is the ratio between the number of hosts and the total number of nodes in the network (in this paper we take $\rho = 0.16$). Hosts are randomly distributed throughout the network.

A routing algorithm is needed to model the dynamic aspects of the network. Packets are created at hosts and sent through the lattice one step at a time until they reach their destination host.

The routing algorithm operates as follows: (1) First a host creates a packet following a distribution defined by the chaotic map (LRD) described above. If a packet is generated it is put at the end of the queue for that host. This is repeated for each host in the lattice. (2) Packets at the head of each queue are picked up and sent to a neighboring node selected according to the following rules: (a) A neighbor closest to the destination node is selected. (b) If more than one neighbor is at the minimum distance from the destination, the link through which the smallest number of packets have been forwarded is selected. (c) If more than one of these links shares the same minimum number of packets forwarded, then a random selection is made.

This process is repeated for each node in the lattice. The whole procedure of packet generation and movement represents one time step of the simulation.

V. Effects on Network Performance of Varying the Underlying Topology

We have compared three different topologies: random, scale-free with $\gamma = 3$, scale-free with $\gamma = 2$, have been compared, while keeping fixed the number of vertices (500) and the edges (3 per node).

In Fig. 1 the number of delivered packets, or throughput, has been plotted as a function of the generation rate, $\lambda$, for the three considered topologies. Scale-free networks show a vanishing value of the critical load $\lambda$, i.e. the value of $\lambda$ at which the networks under high traffic rates. It is worth noting that this is in strong agreement with results shown in [12].

Notice that the differences among the different considered topologies, increase for higher values of $\lambda$. In particular random networks seem to behave better than other networks under high traffic rates. It is worth noting that this is in strong agreement with results shown in [12].

In Fig. 2, the delivery time for packets to reach their destination has been plotted versus the generation rate, $\lambda$. The results are in accordance with those for throughput: the highest delivery time have been achieved for random networks, the lowest for scale-free networks with $\gamma = 2$.

The reason for this is that packets that are stored in the routers’ queues without being delivered to their destination, increase the time needed for other packets to reach their destination. Moreover scale-free networks show a vanishing value of the critical load $\lambda$, i.e. the value of $\lambda$ at which the networks under high traffic rates. It is worth noting that this is in strong agreement with results shown in [12].

In Fig. 2, the delivery times versus the generation rate, $\lambda$. Three different networks, random, scale-free with $\gamma = 3$, scale-free with $\gamma = 2$, have been compared, while keeping fixed the number of vertices (500) and the edges (3 per node).
which a phase-transition occurs [4], with respect to random graphs. Consequently, although scale-free networks are characterized by a shorter characteristic-path-length [13], they show worst performances in terms of communication.

It is somewhat surprising that the structure of scale-free graphs, which are ubiquitous in nature, does not lead to any benefit but rather a worsening in terms of the end-to-end performance. In particular, the characteristic parameters known as throughput and delivery time are considerably affected by the congestion at the network hubs. This is counter-intuitive when one considers that the shortening of the distances in the network might result in a reduction of the delivery time and thus an increase of the throughput.

This interesting phenomenon is analogous to the paradox of heterogeneity [14], which has been observed in the context of synchronizability of scale-free networks.

VI. TOPOLOGY-AWARE COMMUNICATION MODELS

Sofar we have considered no differences in the communication behavior among nodes characterized by different degree. Yet, it is unrealistic to assume that resources such as bandwidth are uniformly distributed among the network nodes in strongly heterogenous networks. Instead, it is very likely that hubs, which are characterized by a high number of incoming and outgoing links, are found to play a fundamental role in communication over the network. They are typically characterized by having higher server strength transmission rates and larger buffers than more peripheral nodes.

For these reasons, in [? we have introduced the following topology-aware communication model: (i) the transmission rate \( r \) is assumed to scale with the degree at each vertex \( i, k(i) \), as: \( r(i) = c_1 k(i)^{\alpha} \) (note that in the particular case where \( \alpha = 0 \), we recover the original case, with all the nodes having the same transmission rates); (ii) the maximum queue length (i.e., the buffer size) is no longer assumed infinite but is taken to scale with the degree at each vertex \( i, k(i) \), as: \( q(i) = c_2 k(i)^{\beta} \). In what follows, we analyze separately, by means of numerical simulations, the effects of varying \( \alpha \) and \( \beta \) on the network communication performance. As a representative case, we assume \( c_1 = 1 \) and \( c_2 = 50 \). Similar behavior was observed for other values of \( c_1 \) and \( c_2 \).

The routing algorithm is the same as the one described above, with the difference that at each iteration of the algorithm, a third step is considered. Namely, packets at the head of each queue, exceeding its maximum capability, are dropped.

VII. NETWORK PERFORMANCE

Using the network model and traffic generator detailed above, simulations were carried out to analyse various aspects of end-to-end performance for two types of network. Namely, results for random graphs have been paired with those of scale-free graphs with \( \gamma = 3 \). We have calculated the corresponding output for scale-free graphs with \( \gamma = 2 \) and have found that the differences in behaviour with the alternative value \( \gamma = 3 \) are negligible by comparison with the behaviour of the random graph, and so the third set of comparisons is not repeated here.

In Fig. 3 we see that random graphs respond more quickly with smaller delivery times as \( \alpha \) increases (from zero). Similarly, it is observed the communication is much more efficient in terms of delivered packets at high loads (or generation rate) as \( \alpha \) increases. The number of delivered packets, instead, is observed to be unaffected by the buffer sizes at the nodes, being mainly determined by the network topology. Finally, in Fig. 4, the number of dropped packets is observed to decrease as the buffer sizes are scaled more sensitively with vertex degree. (More evidence can be found in [1],[?] where further simulation results are reported.)

VIII. CONCLUSIONS

We have shown how topological transitions in a given network from random to scale free affect the load distribution on the network itself. In particular, we characterised such load distribution in terms of the average load and its standard deviation. We observed that as the topological transition takes place, the network performance worsens and the load tends to become more localised (higher standard deviation). Moreover, by introducing a topology-aware communication model, we showed how it is possible to counter-balance some undesirable effects due to the topology.

REFERENCES

Fig. 3. Delivery time versus the generation rate, $\lambda$. The network is a (a) random graph ($\gamma = \infty$), (b) scale-free graph ($\gamma = 3$) with number of nodes $N = 512$ and number of edges $M = 2N$. We show the effects of varying the transmission rates $r(i)$ at node $i$, according to the law $r(i) = k(i)^\alpha$, for $\alpha$ ranging between 0 and 0.5 (blue to red). The black dotted line represents the free regime at $\alpha = 1$. 

Fig. 4. Number of dropped packets versus the generation rate, $\lambda$. The considered network is scale-free with $\gamma = 3$ and $\gamma = \infty$, the number of nodes being $N = 512$ and number of edges $M = 2N$. We show the effects of varying the queue length $q(i)$ at node $i$, according to the law $q(i) = 50k(i)^\beta$, for $\beta$ ranging between 0 and 1.5.
Queuing Networks with Periodic Fluid Model

Tommaso Addabbo† and Ljupco Kocarev‡

†Information Engineering Department, University of Siena
Via Roma, 56, Siena 53100, Italy
‡Institute for Nonlinear Science, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0402
Email: addabbo@dii.unisi.it, lkocarev@ucsd.edu

Abstract—In this paper we consider state-dependent Markovian queuing networks and we provide a formal definition of their associated fluid model. We particularly consider queuing networks whose fluid models have global periodic attractors, focusing on the relationships which may occur in such cases between the network stochastic dynamics and the deterministic dynamics of its corresponding fluid model.

1. Introduction

Queuing networks has been successfully utilized for the analysis of several different real-world problems. In the parlance of queuing theory, a queuing system is a place where customers arrive to receive service from a service facility (i.e., a server). In these models the generic terms “customers” and “servers” may represent various entities depending on the application. In the case of a data switching network, “customers” are packets that arrive at the switching node and “servers” are transmission channels. Recently, queuing theory has been considered for the modeling and understanding of the joint action of genes, proteins and RNA molecules (interwoven in intricate interdependencies commonly known as genetic networks) [1]. In this approach, customers represent molecules of some sort, waiting for chemical reactions.

In this paper we consider Markovian (or exponential) queueing networks, for which arrival and service rates, as well as routing probabilities, may depend on queue lengths. Following Mandelbaum and Pats [2], for a such given queuing network we provide a formal definition of its associated fluid model, and we focus on the relationships which may occur between the network stochastic dynamics and the deterministic dynamics of its corresponding fluid model. We give an example of queuing network whose fluid model has a globally attractive limit cycle and as an interesting issue we show that, depending on the network parameters, the fluid model allow for a statistical characterization of the network long-run dynamical behavior. The example is properly chosen to show that the analysis of the fluid model makes sense also when some components of the queuing network periodically become empty (weak load condition).

2. The stochastic model

We consider an n-node open queuing network in which each node is a $M_i/E_1/1 - FIFO$ single station with one server [1]. We denote with $Q_i(t)$ the total number of customers at node $i$ at time $t$ (i.e., the number of customers in service plus the number of customers waiting in the buffer $i$), while $\lambda_i$ and $\mu_i$ denote the exogenous arrival rate and service rate for the node $i$, respectively. After being served at node $i$, a customer can randomly either leave the system or instead he can join any queue. In the latter case, the customer who is just leaving node $i$ will join the queue at node $j$ depending on the transition probability $P_{ij}$ which characterizes the customer flow from node $i$ to node $j$. The $n \times n$ square matrix $P = (P_{ij})$ is the transition probability matrix of the queuing network, where $0 \leq P_{ij} \leq 1$ and $\sum_{j=1}^{n} P_{ij} \leq 1$. In this paper we assume that both arrival rates and service rates as well as the transition matrix are locally Lipschitz continuous functions of the state $Q$.

In this case, by denoting with $\mathbb{N}$ the set of natural numbers, the dynamic evolution of the queuing network is the realization of a non-negative valued stochastic process $Q = (Q(t), t \geq 0)$ that satisfies a relation of the form

$$Q(t) = Q(0) + A(t) + F(t) - D(t), \quad t \geq 0,$$

where the column vector $Q(0)$ is the initial state of the queuing network, $A(\cdot), F(\cdot), D(\cdot)$ are $n$-dimensional counting processes related to the customer arrivals from the outside of the network (exogenous arrivals), to the customer flows inside the network (endogenous arrivals), and to the service completions (departures), respectively. For the $i$-th component in (1) we assume

$$A_i(t) = N_{i} \left( \int_{0}^{t} \lambda_i(Q(s)) ds \right)$$

$$D_i(t) = N_{i} \left( \int_{0}^{t} \mu_i(Q(s)) 1_{[0,\infty)}(Q(s)) ds \right)$$

$$F_i(t) = \sum_{j=1}^{n} \int_{0}^{t} 1_{s_i(s-)} \left( U_j D_j(s) dD_j(s) \right)$$

where in general the writing $Q_i(s-)$ denotes the left limit of $Q_i(s)$ at $s$, assuming $Q_i(0-)=Q_i(0)$. In the above equations the used notation has the following meaning: for
i = 1, ..., n, the quantities $N_i^+(\cdot), N_i^-(\cdot)$ are mutually independent standard Poisson processes (with rate 1). Given any set $B$, function $1_B : B \to \{0, 1\}$ is an indicator function defined as $1_B(x) = 1$ if $x \in B$, or $1_B(x) = 0$ otherwise. In (3) the set $(0, \infty)$ is the set of strictly positive reals and the indicator function $1_{(0,\infty)}(\cdot)$ is used since no customer departure from a station can occur while this station is empty. For $j = 1, ..., n$, the sequences $\{U_j(t)\}_{t \in \mathbb{R}^+}$ are assumed to be sequences of i.i.d. random variables uniformly distributed on $[0, 1]$, and for $i, j = 1, ..., n$ and $s \geq 0$, the set $\pi_{ji}(s-) = \{ x \in [0, 1] : \sum_{k=1}^n P_{jk}(Q(s)) \leq x < \sum_{k=1}^n P_{jk}(Q(s-)) \}$, where for $j, k = 1, ..., n$, the terms $P_{jk}$ are elements of the transition matrix function $P$. In this expression, the left limit is used since, when a customer is served at time $t_0$, she/he is instantaneously routed according to the transition matrix $P(Q(t_0-))$ which refers to the networks state at $Q(t_0-)$.

Indeed, the transition matrix $P(Q(h))$ is defined only when the customer served at $t_0$ has been definitely routed somewhere. The introduced above quantities have the following interpretations: for $t \geq 0$, $A_i(t)$ and $F_i(t)$ represent the cumulative number of exogeneous and endogenous arrivals respectively to station $i$ during $[0, t]$; when a customer at time $t$ is served at node $j$, the quantity $D_i(t)$ increases by 1 from $D_i(t-)$. If the random variable $U_j(D_i(t)) \in \pi_{ji}(t)$ then the customer departing form node $j$ is routed to node $k$. Note that for each $t \geq 0$, $\bigcup_{i=1}^n \pi_{ji}(t) \subseteq [0, 1]$. It is worth emphasizing that we assume the processes $N_i^+$ and $N_i^-$ and the sequences $\{U_j(t)\}_{t \in \mathbb{R}^+}$ all mutually independent. Moreover we assume that the spectral radius, $r(P(\cdot))$, of the transition matrix $P$ such that $r(P(\cdot)) < 1$ for all $\nu \in \mathbb{R}^n$ and that the two vector-valued functions $\lambda$ and $\mu$ satisfy a linear growth constraint, that is there exists a positive constant $A$ such that for all $\nu \in \mathbb{R}^n$, it results $\max(||\lambda(\nu)||, ||\mu(\nu)||) \leq A(1 + ||\nu||)$. As discussed in [2], since the queuing network process described above is a Markov jump process on $\mathbb{N}^n$, this latter assumption represents a sufficient condition for the non-explosion of the network’s state $Q$, that is $P(||Q(t)|| < \infty, t \geq 0) = 1$. Almost surely, the paths of the stochastic process $\{Q(t), t \geq 0\}$ consist of $n$ piecewise constant right-continuous non-negative integer-valued functions with finite left limits. Moreover, for $i = 1, ..., n$, stochastic processes $\{A_i(t), t \geq 0\}$, $\{D_i(t), t \geq 0\}$ can be referred to as time-changed Poisson processes.

3. Fluid Models and Fluid Approximations

Fluid approximations are formal deterministic approximations to queuing networks. Frequently these are obtained as scaling limits, where arrival and service rates are speeded up, while attenuating the stochastic fluctuations by a same quantity. The limit for this procedure suppresses the random fluctuations around what appears to be the main trajectory of the process. From a theoretical point of view, a fluid model of a queuing network is obtained as the limit of a sequence of stochastic processes $q^i(t) = [q^i(t), t \geq 0]$ which is determined by means of simple transformations of parameters of the starting stochastic system [1]. Here, the superscript $k$ indicates that the corresponding quantity is related to the $k$-th network of such a sequence. Accordingly, let us consider a queuing network with initial condition $Q_0$ and whose stochastic evolution is described by a random process $Q$ satisfying relationship (1) and arising assumptions. We introduce with $k = 1, 2, \ldots$ a sequence of stochastic processes $q^i_k$, all of them having the same initial condition $Q_0$, such that

$$q^i_k = Q_0$$

$$A^i_k(t) = k\lambda(v)$$

$$\mu^i_k(v) = k\mu(v)$$

$$P^i_k(v) = P(v)$$

where $v \in \mathbb{R}^n$ and such that for each $k \in \mathbb{N}$ the stochastic process $q^i_k$ satisfies relationship (1) and subsequent assumptions, i.e.

$$q^i(t) = Q_0 + A^i(t) + F^i(t) - D^i(t) \quad t \geq 0$$

By construction, in the Markovian stochastic process described by $q^i_k$ an arrival event in node $i$ is recorded by increasing the length of the $i$th corresponding rational-valued “queue” by the quantity $1/k$, and the same quantity is subtracted when a departure occurs. Accordingly, with reference to the queuing network described by $Q$, the process $q^i_k$ has arrival and service rates speeded up $k$ times to compensate for the $k$-times reduced perturbation effects on the state by the stochastic fluctuations. Note, also, that if $k = 1$ then $q^i = Q$, i.e. the stochastic process $Q$ describing the original queuing network agrees with the first element of the sequence $q^i_k$ of rescaled stochastic processes.

For the dynamical behavior of the asymptotic process $q^i_k$ when $k \rightarrow \infty$ the following theorem holds.

Theorem 1 (Fluid Limit [2]) The sequence of stochastic processes $q^i(t), t \geq 0$ defined in (5) - (10) for $k \rightarrow \infty$ converges in probability uniformly over compacts in $[0, \infty]$ to the unique non-negative absolutely continuous function $q : \mathbb{R}_+ \rightarrow \mathbb{R}^n$ which is for $t \geq 0$ the solution of

$$q(t) = Q_0 + \int_0^t [\lambda(q(s)) + [P(q(s)) - I][\mu(q(s))] ds + r(t),$$

664
where the term \( r(t) \) takes into account the cumulative effect of the system’s state-dependent reflection on the boundaries

\[
r(t) = \int_0^t [1 - P'(q(s))] \, dy(s), \quad t \geq 0,
\]

where the function \( y : \mathbb{R}_+ \rightarrow \mathbb{R}_+^n \) is continuous, non-negative, and non-decreasing components, such that \( y(0) = 0 \) and

\[
\int_0^\infty I_{(0,\infty)}(q(t)) \, dy(t) = 0.
\]

The unique solution \( q \) of (11) is referred to as the fluid limit of the sequence of processes \( \{q^t(t), t \geq 0\} \) derived from the stochastic process \( \{Q(t), t \geq 0\} \) with initial condition \( Q_0 \). By condition (13), since \( y \) is non-negative and due to the continuity of \( q \), it turns out that if the initial condition \( Q_0 \) does not have any components on the boundaries, then there exists an interval \([0, t_1]\) for which \( r(t) = 0 \) \( \forall t \in [0, t_1] \). Under this hypothesis, the fluid limit \( q \) over \([0, t_1]\) agrees with the unique solution of the system of ordinary differential equations (ODEs)

\[
\frac{dq}{dt} = \lambda(q) + [P'(q) - I] \mu(q) \equiv h(q),
\]

when assuming \( q(0) = Q_0 \). In general, when taking into account the effects of the boundaries, instead of (14) the following equation must be considered [2]:

\[
\frac{dq}{dt} = h(q) + [I - P'(q)] \hat{m}(q) \equiv \hat{H}(q),
\]

where the function \( h \) is defined in (14) and function \( \hat{m} \) is unambiguously determined by the state-dependent reflection, as described in the following. Referring to the introduced notion, let denote by \( I(t) \subseteq \{1, 2, \ldots, n\} \) the subset of indices \( i \) such that \( h_i(q) < 0 \) and such that the corresponding state’s component \( q_i \) is zero, i.e. we have \( i \notin I(q) \iff q_i = 0, h_i(q) < 0 \), for \( i = 1, \ldots, n \). Accordingly, the components of \( \hat{m}(q) \) can be obtained by solving the linear system:

\[
\begin{align*}
  h_q(q) + \hat{m}(q) &= \sum_{j \in I(q)} P_{ji}(q) \hat{m}_j(q) = 0, & \text{if } i \in I(q), \\
  \hat{m}_i(q) &= 0, & \text{otherwise}
\end{align*}
\]

Expression (15) is referred to as the fluid model equation for the queuing network with queue length process \( Q \).

### 3.1. A Periodic Fluid Model Example

In a state dependent queuing network, since no strong constraints are imposed on the resulting functional form of the deterministic fluid model, the resulting dynamical system (15) can display a wide range of dynamics, including steady states, periodic behavior and chaos. To give an example, let us consider a simple two-station \( M_\infty/M_\infty/1 -

**FIFO network with parameters**

\[
\lambda_0(Q_E) = \begin{cases} 6E, & \text{if } 0 \leq q \leq E \leq 50, \\
6E + 20f'E, & \text{if } q > E > 50.
\end{cases}
\]

\[
\mu_\infty(Q_E) = \begin{cases} 3E, & \text{if } 0 \leq q \leq E \leq 500, \\
3E + 20f'E + 4f^2E, & \text{if } q > E > 500.
\end{cases}
\]

\[
\lambda^E = 0, \quad \mu^E = 5E, \quad P = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.
\]

where \( E > 0 \), \( f^E = Q^E - E \), \( f^E = Q^E - E \) and, for any real number \( n \), we have \( (a)^n = \max(0, a) \). By changing parameter \( E \) we change the way the network reacts to its workload, and for each \( E \in \mathbb{R}^+ \) the stochastic process \( Q^E \), describing the queuing network with parameters (17)-(19), satisfies all assumptions discussed in Section 2. Therefore, referring to (15), the fluid model of \( Q^E \) has the form

\[
\begin{align*}
 h^E(q) &= \begin{cases} 3E, & \text{if } 0 \leq q \leq E \leq 50, \\
3E - 4gE, & \text{if } q > E > 50,
\end{cases} \\
 h^E(q) &= \begin{cases} -2E, & \text{if } 0 \leq q \leq E \leq 50, \\
-2E + 20gE, & \text{if } q > E > 50,
\end{cases}
\end{align*}
\]

where \( g = q - E \) and \( g = q - E \). The analysis of (20)-(21) reveals that there are points on the boundary \( (q_\ast = 0) \) in which \( h^E \) is negative for all \( E > 0 \), and that there are some points on the boundary \( (q_\ast = 0) \) in which \( h^E \) is negative for all \( E > 0 \). Indeed it results \( h^E(0, q_\ast) < 0 \) if \( q > E/3 \) and \( h^E(q_\ast, 0) < 0 \) if \( 0 \leq q_\ast < 11E/10 \). Accordingly, solving system (16), function \( \hat{m}^E \) has components

\[
\begin{align*}
 \hat{m}_1^E(q_\ast) &= \begin{cases} -h_1(q_\ast) = 9q_\ast - 12E, & \text{if } q_\ast = 0, g^E > 4E/3, \\
0, & \text{otherwise},
\end{cases} \\
 \hat{m}_1^E(q_\ast) &= \begin{cases} -h_1(q_\ast) = 22E - 20q_\ast, & \text{if } 0 \leq q_\ast \leq 11E/10, \\
0, & \text{otherwise}.
\end{cases}
\end{align*}
\]

The dynamical system (15) with \( h^E \) and \( \hat{m}^E \) defined as in (20)-(21) and (22)-(23) has the global periodic attractor plotted in solid line in Fig. 1.b for \( E = 500 \). A rigorous analysis of the global attractiveness of the limit cycle is here omitted, however it is not difficult to verify that for this system every trajectory must eventually hit the boundary \( q_\ast = 0 \) in \( q_\ast > E/3 \) entering in a finite time the periodic cycle after a transient evolution flattened on the \( q_\ast \) axis. In the same figure, two stochastic dynamics of two queuing networks with primitives (17)-(19) for \( E = 50 \) and 500 respectively are plotted.
Figure 1: (a): dynamical evolution of \((q_x,q_y)\) for the fluid trajectory over the limit cycle \((E = 500)\); (b): the global attractiveness of the limit cycle is here shown for several initial conditions \((E = 500)\). (c) and (d): stochastic dynamics of two queuing networks with primitives (17)-(19) for \(E = 50\) and \(E = 500\) respectively.

4. Periodic Queuing Networks: Long-Run Behavior

In this section we analyze the stochastic dynamics from a probabilistic point of view. In particular, we show that for queuing networks with periodic fluid models the system state under the stochastic evolution approaches a stationary invariant distribution which is intimately related to the unique invariant probability measure existing for the periodic fluid model [1]. To get insight this point, let write (15) as a dynamical system described by \(\frac{dq}{dt} = H(q)\), and let assume the existence of a globally attractive limit cycle with support \(\varphi \subseteq \mathbb{R}^n\). By denoting with \(\mathcal{B}(\mathbb{R}^n)\) the standard Borel \(\sigma\)-algebra over \(\mathbb{R}^n\), it can be proved [1] that the function \(\Pi : \mathcal{B}(\mathbb{R}^n) \rightarrow [0,1]\) defined by the curvilinear integral \(\Pi(A) = \frac{1}{q_0} \int_{q_0}^A H(q) \, dq\) is a probability measure over \(\mathbb{R}^n\). The following property holds [1]

**Theorem 2** The measure \(\Pi\) is the unique invariant probability measure over \(\mathbb{R}^n\) with respect to the periodic dynamical system.

In eventually periodic systems the stationary distribution related to the probability measure \(\Pi\) is impulsive over the zero-measure support of the limit cycle. In such case, the unique stationary (impulsive) distribution is described over \(\varphi\) by a curvilinear probability density function (pdf) \(\psi : \varphi \rightarrow \mathbb{R}_+\), defined as \(\psi(q) = \frac{1}{\sqrt{2\pi}}\).

We now provide some numerical evidence that the state distribution of a queuing network with periodic fluid model approaches under the evolution a stationary invariant distribution strongly related to the unique impulsive stationary distribution of the periodic fluid model. Referring to our example, this issue appears clear in Fig. 2. The numerical analysis of the stochastic evolution showed that, regardless of the the initial distribution, a crater-shaped stationary distribution is approached, being the ridge of the crater positioned over the support \(\varphi\) of the fluid limit cycle. This result can be explained observing that the dynamics in queuing networks evolves under the strong influence of the underlying fluid deterministic velocity field. Since in our case the fluid system has a globally attractive limit cycle, over the periodic trajectory the deterministic state variable which corresponds to the zero Lyapunov exponent, the phase variable, is neutrally stable and hence, depending on the system parameters, the dominant effect of the stochastic perturbations is a phase diffusion, and around \(\varphi\) a distribution proportional to the natural invariant curvilinear pdf \(\psi\) is consequently approached.

In conclusion, the dynamical analysis of the fluid model can provide useful information for predicting the evolution of a queuing network, especially when its fluid model has a globally attractive limit cycle. The example has shown that the queuing network dynamics can be strongly influenced by the deterministic velocity field even if the queues become empty (i.e. even if the network has a small workload). This is in sharp contrast to the intuitive idea that fluid approximations can be sensibly taken into account only for queuing networks with overloaded servers.

**References**


Adaptive traffic in complex networks
A. Fekete†, G. Vattay† and L. Kocarev‡
†Collegium Budapest, Institute for Advanced Studies, Szentháromság u. 2., H-1014 Budapest, Hungary
‡Institute for Nonlinear Science, University of California San Diego, La Jolla (CA), USA,
Macedonian Academy of Sciences and Arts, Skopje, Macedonia
Email: fekete@complex.elte.hu, vattay@elte.hu, lkocarev@ucsd.edu

Abstract—We present a method for estimating the rates of flows across an arbitrary network, determined by the TCP flow control. The method is simple, fast and can estimate the performance of a TCP network.

1. Introduction

Traditional approaches to performance evaluation packet networks have normally relied on attempts to describe as closely as possible the dynamics of network elements over a discrete state-space. However, discrete models, requiring the description of the dynamics of the different network elements over their discrete state spaces, suffer from limited scalability, thus allowing only the performance analysis of rather small networking setups. A new class of semi-analytical models has recently been introduced in the networking arena, and today appears to be the most promising approach for scalable and accurate performance analysis of large IP networks. This new approach, that is often called ‘fluid models’, adopts an abstract deterministic description of the average network dynamics through a set of ordinary differential equations [1, 2, 3, 4, 5], thus neglecting the short term, packet-by-packet description of the stochastic network dynamics. The majority of traffic on the current Internet uses Transmission Control Protocol (TCP) based transport. A critical component of the TCP is the dynamic window flow control designed to prevent congestion collapses. In this paper we develop models to predict network performance generalizing the work of [6, 7]. The methodology developed here can be used to solve a number of engineering problems, such as estimating the capacity required in the network to meet certain performance objectives, for example providing a minimum throughput per source.

2. Network Model

We model a computer network by graph \( G = (V, E) \), where \( V \) denotes the set of nodes (computers, routers, switches, etc.), and \( E \) denotes the set of directed links between nodes. The capacity of link \( e \in E \) is denoted by \( C_e \). If an incoming packet finds the link busy, then the packet is stored in a buffer of size \( K_e \), supposed the buffer is not full. If the buffer is full, then the incoming packets are dropped.

Data packets are transmitted from a source node to a destination node through intermediate buffers. The route of a flow is a link sequence \( f = (e_1, e_2, \ldots, e_f) \in F \), where \( F \) denotes the set of flows in the network, and \( l_f \) denotes the length of flow \( f \). In a computer network packets are usually routed via (one of) the shortest path(s) between the source and destination nodes. In our model we allow arbitrary simple path as a route, that is any route without repeated vertices.

Let us define the routing matrix \( I : E \times F \rightarrow \{0, 1\} \) as:

\[
(e, f) \mapsto I_{e,f} = \begin{cases} 1, & \text{if } e \in f, \\ 0, & \text{if } e \notin f. \end{cases}
\]

(1)

It is obvious that the load of link \( e \) can be given by

\[
L_e = \sum_{f \in F} I_{e,f} \tag{2}
\]

and the length of flow \( f \) is

\[
l_f = \sum_{e \in E} I_{e,f} \tag{3}
\]

For any \( k \in \mathbb{N} \) let us define the following decomposition \( I^{(k)} : E \times F \rightarrow \{0, 1\} \) of the routing matrix as well:

\[
(e, f) \mapsto I^{(k)}_{e,f} = \begin{cases} 1, & \text{if } e = e_k \in f, \\ 0, & \text{if } e \neq e_k \in f, \end{cases}
\]

(4)

where \( e_k \) denotes the \( k \)th element of the route, that is the matrix contains only the \( k \)th element of the routes. Is easy to show that \( I = \sum_{k=1}^{\infty} I^{(k)} \).

Let us denote the traffic intensity of flow \( f \) before buffer \( e \) by \( X_{e,f}^{in} \). Due to packet losses at buffer the traffic intensity of the flow \( f \) after buffer \( e \) will change as:

\[
X_{e,f}^{out} = F(X_{e,f}^{in}, X_{<e}^{in}; K_e, C_e), \tag{5}
\]

where \( X_{<e} \) denotes the aggregated traffic at buffer \( e \):

\[
X_{<e}^{in} = \sum_{f \in F} I_{e,f} X_{e,f}^{in} \tag{6}
\]
where $S_e$ denotes the source intensity at buffer $e$. The particular form of function $F$ depends on the queuing model of the buffers.

The transfer probability of flow $f$ at buffer $e$ is obviously:

$$p^+_e = 1 - \frac{X_{e,f}^{\text{in}} - X_{e,f}^{\text{out}}}{X_{e,f}^{\text{in}}} = \frac{F(X_{e,f}^{\text{in}}; X_{e,f}^{\text{in}}; K_e, C_e)}{X_{e,f}^{\text{in}}}. \quad (7)$$

Let us suppose that at a buffer the transfer probability is independent of which flow the packet belongs to, that is $p^+_e \equiv p^+_r$. It follows that the right hand side of Eq. (7) is independent of the individual flows as well:

$$\frac{F(X_{e,f}^{\text{in}}; X_{e,f}^{\text{in}}; K_e, C_e)}{X_{e,f}^{\text{in}}} \equiv G(X_{e,f}^{\text{in}}; K_e, C_e). \quad (8)$$

Therefore, the transfer probability $p^+_r$ at buffer $e$ is:

$$p^+_e = G(X_{e,f}^{\text{in}}; K_e, C_e), \quad (9)$$

and the transfer probability experienced by flow $f$ can be given as:

$$p^+_f = \prod_{e \in f} p^+_e = \prod_{e \in f} G(X_{e,f}^{\text{in}}; K_e, C_e). \quad (10)$$

With the help of the routing matrix the above probability can be written as:

$$p^+_f = \prod_{e \in E} \left[ G(X_{e,f}^{\text{in}}; K_e, C_e) \right]^{I_{e,f}} = \exp \left[ \sum_{e \in E} I_{e,f} \ln \left( G(X_{e,f}^{\text{in}}; K_e, C_e) \right) \right]$$

The intensity of the source of flow $f$ can be written as the function of the above transfer probability as:

$$S_f = S(p^+_f). \quad (11)$$

For example, for a non-elastic traffic $S(p_f) \equiv \text{const}$. For a TCP connection, however, the flow intensity can be approximated by the so called “inverse square-root formula”:

$$S(p^+_f) = \frac{A}{\sqrt{1 - p^+_f}} \quad (12)$$

Let us investigate Eq. (6) now. With the decomposition of the routing matrix one can write

$$X_{e,f}^{\text{in}} = I_{e,f}^{(1)} S_f + I_{e,f}^{(2)} \left( \sum_{e_i \in E} I_{e_i,f}^{(1)} G_{e_i} \right) S_f + \cdots + I_{e,f}^{(k-1)} \left( \sum_{e_i \in E} I_{e_i,f}^{(1)} G_{e_i} \right) S_f + \cdots, \quad (13)$$

where the shorter $G_e \equiv G(X_{e,f}^{\text{in}}; K_e, C_e)$ notation has been used. That follows that the following transfer matrix describes the traffic flows in the network:

$$T: ExF; \quad (e,f) \mapsto T_{e,f} = \sum_{k=1}^{\infty} I_{e,f}^{(k)} \prod_{e_i \in E} \left( \sum_{e_i \in E} I_{e_i,f}^{(1)} G_{e_i} \right) \quad (14)$$

For an M/M/1/K queue the form of $G_{e} \equiv G(X_{e,f}^{\text{in}}; K_e, C_e)$ can be given as follows. The probability of forming a queue of length $k$ is

$$P(Q_e = k) = \frac{1 - \rho_e}{1 - \rho_e K_e + \rho_e p_e}, \quad 0 \leq k \leq K_e, \quad \rho_e = \frac{X_{e,f}^{\text{in}}}{C_e} \quad (15)$$

where $\rho_e = X_{e,f}^{\text{in}}/C_e$ denotes the utilization of link $e$. The probability that link $e$ is blocked equals to the probability that the queue length is $K_e$. Therefore

$$G(X_{e,f}^{\text{in}}; K_e, C_e) = 1 - P_e(K) = 1 - \frac{1 - \left( X_{e,f}^{\text{in}}/C_e \right)^{K_e}}{1 - \left( X_{e,f}^{\text{in}}/C_e \right)^{K_e+1}} \quad (16)$$

In the limit of infinite buffer, which consists of the fluid approximation of the traffic, the form of $G(X_{e,f}^{\text{in}}; K_e, C_e)$ is

$$G_{\infty}(X_{e,f}^{\text{in}}; C_e) \equiv \lim_{K_e \to \infty} G(X_{e,f}^{\text{in}}; K_e, C_e) = \min \left( 1, \frac{C_e}{X_{e,f}^{\text{in}}} \right) \quad (17)$$

For the average queue length let us calculate the Z-transform of Eq. (15):

$$\phi_{Q_e}(z) = \sum_{k=0}^{\infty} P(Q_e = k) z^k = \frac{1 - \rho_e}{1 - \rho_e K_e + \rho_e p_e} = \frac{1 - \rho_e}{1 - \rho_e K_e + \rho_e p_e} \quad (18)$$
3. Distribution of the Cumulative Queue Length

In this section we calculate the distribution of the cumulative queue length \( Q_f = \sum_{e \in f} Q_e \) experienced by flow \( f \in F \). If we assume that the queues are independent, then the Z-transform of the cumulative queue length can be given as follows:

\[
\phi_{Q_f}(z) = \prod_{e \in f} \phi_{Q_e}(z) = \prod_{e \in f} \left[ \frac{1 - \rho_e}{1 - \rho_e z} \right]^{K_e + 1}\]

(20)

The partial fraction expansion of the first denominator yields that

\[
\prod_{e \in f} \frac{1}{1 - \rho_e z} = \sum_{e \in f} \frac{1}{1 - \rho_e z} \prod_{e' \in f, e' \neq e} \frac{\rho_e^{n_f}}{\rho_e - \rho_{e'}}
\]

\[
= \sum_{n=0}^{\infty} \sum_{e \in f} \prod_{e' \in f, e' \neq e} (\rho_e - \rho_{e'})^{-n} z^n, \quad (21)
\]

where \( \rho_e \neq \rho_{e'} \) for all \( e \neq e' \in f \) and \( n_f \) denotes the length of flow \( f \), that is the number of its elements. The inverse transform of Eq. (21) can be obtained from the coefficients of \( z^n \):

\[
F(n) = \sum_{e \in f} \prod_{e' \in f, e' \neq e} \frac{\rho_e^{n_f + n - 1}}{\rho_e - \rho_{e'}}, \quad \text{if } n \geq 0, \quad (22)
\]

and \( F(n) \equiv 0 \) if \( n < 0 \). If \( \rho_{e_1} = \rho_{e_2} = \cdots = \rho_{e_k} \equiv \rho \) for some \( e_1 \neq e_2 \neq \cdots \neq e_k \in f \) then one should take the \( \rho_{e_1} \to \rho, \rho_{e_2} \to \rho, \ldots, \rho_{e_k} \to \rho \) limits. In the special case \( \rho_e \equiv \rho \) for all \( e \in f \):

\[
F(n) = \left( \frac{n + n_f - 1}{n} \right) \rho^n, \quad \text{if } n \geq 0 \quad (23)
\]

and \( F(n) \equiv 0 \) if \( n < 0 \).

To evaluate the inverse Z-transformation on (20), we need to expand the nominator of the second fraction of Eq. (20) into power series as well:

\[
\prod_{e \in f} \left[ 1 - (\rho_e z)^{K_e + 1} \right] = \sum_{k=0}^{n_f} \sum_{s \subseteq f} (-1)^k \left[ \prod_{e \in s} \rho_e^{K_e + 1} \right] z^k + \sum_{s \subseteq f} K_s (24)
\]

Since the multiplication with \( z^k \) on the Z-transform domain corresponds to an argumentum shift on the normal domain, therefore the probability distribution can be given by:

\[
P(Q_f = n) = \left[ \prod_{e \in f} \frac{1 - \rho_e}{1 - \rho_e z} \right]^{n_f} \sum_{k=0}^{n_f} \sum_{s \subseteq f} (-1)^k \left[ \prod_{e \in s} \rho_e^{K_e + 1} \right] F \left( n - k - \sum_{e \in s} K_e \right) \quad (25)
\]

3.1. Infinite buffers

Let us suppose that \( K_e = \infty \) for all \( e \in f \). Then only one term of (25) does not vanish:

\[
P(Q_f = n) = F(n) \prod_{e \in f} (1 - \rho_e) \quad (26)
\]

3.2. Finite, identical buffers

Let us suppose that buffers are finite, but the same size: \( K_e = K \) for all \( e \in f \). Then:

\[
P(Q_f = n) = \prod_{e \in f} \frac{1 - \rho_e}{1 - \rho_e z}^{n_f} \sum_{k=0}^{n_f} (-1)^k \left[ \prod_{e \in s} \rho_e^{K_e + 1} \right] F \left( n - k (K + 1) \right) \quad (27)
\]
3.3. Finite identical buffers and identical utilizations

Let us suppose that $K_e \equiv K$ and $\rho_e \equiv \rho$ for all $e \in f$. Then

$$P(Q_f = n) = \left(1 - \frac{\rho}{1 - \rho^{K+1}}\right)^{n_f} \sum_{k=0}^{n_f} (-1)^k \binom{n_f}{k} \left(\frac{n + n_f - k(K + 1) - 1}{n_f - 1}\right) \rho^n$$

(28)

4. Numerical Example

As a numerical example, in this section we consider the network shown in Fig. 3. The crucial starting point for the analysis of the network, is (13) which is a fixed point equation for our network model. This equation can be solved using a standard technique such as Newton-Raphson method for example. More detail analysis based on Brouwer’s fixed point theorem can actually ensure that there exists at least one solution of the equation (13). Figure 4 shows the results of our numerical simulations: we show the utilizations of each link versus the buffer size.

References


Information Processing in Networks of Coupled Hindmarsh-Rose Neurons

Paolo Checco†, Marco Righero†, Mario Biey†, and Ljupco Kocarev‡

†Dipartimento di Elettronica, Politecnico di Torino
c.so Duca degli Abruzzi, 24, 10129 Torino, ITALY
‡INLS, University of California, San Diego
9500 Gilman Drive, La Jolla, CA 92093-0402, USA
E-mails: {paolo.checco, marco.righero, mario.biey}@polito.it, lkocarev@ucsd.edu

Abstract—Synchronization plays central role in processing of information in many systems. In this work, synchronization in networks of bursting Hindmarsh-Rose neurons is studied. The influence of the network topology is investigated in detail comparing two different types of coupling: the electrical and the synaptic coupling. In both cases conditions of synchronization with respect to the coupling strength are obtained.

1. Introduction

During the last few years networks of bio-inspired neurons have interested an increasing number of researchers in all branches of science. In particular, spiking neurons have attracted the interest because many studies consider this behavior an essential component in information processing by the brain [1]. In this class of neurons, bursting neurons are of relevant interest because they characterize a variety of biological oscillators. The electrical potential of these neurons, which typically is the state variable that contains the main information, undergoes a succession of alternating active and silent phases in which, respectively, it has a spiking behavior (very fast oscillations) and it evolves slowly without oscillations.

Furthermore, studies suggested that retrieval of the stored patterns is related to spontaneously occurring synchrony in networks of neurons (see [2] for example.) This motivates the investigation of the conditions for synchronization in networks of bursting neurons [3, 4].

The aim of this paper is to investigate the synchronization conditions of a network of Hindmarsh-Rose neurons [5] in two different cases of coupling, namely electrical and synaptic. Many studies have been done in this field [3, 4], but more detailed conditions are here reported. For this purpose, the Master Stability Equation/Function (MSE/F) approach is used [6].

At first, in Sec. 2 some preliminaries are introduced: the Hindmarsh-Rose neuron model, the coupling functions, and the network equations. Then, in Sec. 3, the synchronous state and the MSE and MSF are obtained. In Sec. 4 necessary conditions of synchronization are derived for both cases of coupling and a simple example is shown. Conclusions (Sec. 5) close the paper.

2. Preliminaries

The Hindmarsh-Rose neuron model [5], a simplified version of the Hodgkin-Huxley one and a modification of the FitzHugh equations – originally obtained to model the synchronization of firing of two snail neurons – can be described by the following equations [4, 7], obtained from the original ones by a linear transformation:

\[
\begin{align*}
\dot{x}(t) &= f_x(x, y, z) = ax^2 - y - z \\
\dot{y}(t) &= f_y(x, y, z) = (a + d)x^2 - y \\
\dot{z}(t) &= f_z(x, y, z) = \mu(bx + c - z)
\end{align*}
\]

where \(x(t)\) represents the membrane potential and \(y(t)\) and \(z(t)\) are recovery variables taking account of fast and slow ion currents respectively. Let us use the same parameters as in [4], namely \(a = 2.8, b = 9, c = 5, \alpha = 1.6, \) and \(\mu = 0.001.\) In Fig. 1 the time evolution of the \(x(t)\) variable is depicted, where it is possible to see that, after a short transient, it is periodic.

![Figure 1: Time evolution of the membrane potential \(x(t)\) of a neuron according to the Hindmarsh-Rose model.](image)

It is possible to model the coupling in a network of \(N\) such neurons in different ways. In this work the interest is about the electrical and the synaptic coupling. The synaptic coupling can be modeled as a (static nonlinear) sigmoidal function, which is the simplest one in a neuronal system [8]:

\[
\gamma_s(x_j) = \frac{1}{1 + e^{-\theta_s(x_j - \gamma_s)}}.
\]

As in [4], the free parameters are chosen to be \(\nu = 10\) and \(\theta_s = -0.25.\) On the other hand, the electrical coupling can be modeled with the following linear function:

\[
\gamma_e(x_j, x_i) = (x_j - x_i),
\]
The evolution of the $i$-th neuron is then ruled by
\[
\dot{x}_i(t) = f_x(x_i, y_i, x_i) - g_s \sigma(x_i) \sum_{j=1}^{N} \tilde{c}_{ij} y_j(x_j)
\]
where $\tilde{c}_{ij}$ are the elements of the matrix $\tilde{C}$ and $\tilde{c}_{ii} = 0$, $\tilde{c}_{ij} = \tilde{c}_{ji} = 1$ if neurons $i$ and $j$ are connected to each other, and $\tilde{c}_{ij} = \tilde{c}_{ji} = 0$ otherwise. In the case of electrical coupling $\gamma(x_i, x_j) = \gamma_e(x_i, x_j)$ and $\sigma(x_i) = 1$ and in the case of synaptic coupling $\gamma(x_i, x_j) = \gamma_s(x_i)$ and $\sigma(x_i) = -(x_i - V_s)$.

Let $\xi_i = (x_i, y_i, z_i)$ and $f = (f_x, f_y, f_z)'$, Eq. (4) can be recast as follow:
\[
\dot{\xi}_i = f(\xi_i) + \sigma(\xi_i) \sum_{j=1}^{N} g_{ij} \Gamma(\xi_i, \xi_j)
\]
where $\Gamma(\xi_i, \xi_j) = (\gamma(x_i, x_j), 0, 0)'$, $G = \{g_{ij}\} = g_s C$ is the weighted connectivity matrix, and $C = \{c_{ij}\}$ is the connectivity matrix: in the synaptic case $c_{ij} = \tilde{c}_{ij}$, $\forall i, j$ and in the electrical one $c_{ij} = \tilde{c}_{ij}$ $\forall i \neq j$ and $c_{ii} = -\sum_{j=1}^{N} \tilde{c}_{ij}$.

3. Master Stability Function

In order to obtain the conditions of identical synchronization, the master stability equation/function approach is used [6]. By considering the identical synchronization conditions, $i.e.$ $\xi_1 = \xi_2 = \ldots = \xi_N = \xi^s$, Eq. (5) becomes
\[
\dot{\xi}^s(t) = f(\xi^s) + \sigma(\xi^s) \sum_{j=1}^{N} g_{ij}
\]
It follows that the identically synchronous state exists only if the sum of $g_{ij}$ is constant with respect $i$:
\[
\sum_{j=1}^{N} g_{ij} = g_s \sum_{j=1}^{N} c_{ij} = g_s k.
\]
Note that this condition is automatically respected in the case of electrical coupling because, in this case, $k = 0$.

The evolution of the synchronous state is described by the following system of ordinary differential equation:
\[
\dot{\xi}(t) = f(\xi^s) + k g_s \sigma(\xi^s) \Gamma(\xi^s)
\]
The MSE associated to Eq. (5), true for both electrical and synaptic coupling, is (for details see [6, 9])
\[
\dot{\xi}(t) = \begin{pmatrix}
    Df(\xi^s) + \begin{pmatrix}
        -kg_s \gamma(x^s) & 0 & 0 \\
        0 & 0 & 0 \\
        0 & 0 & 0
    \end{pmatrix} + \\
    (\sigma(\alpha) \gamma'(x^s) & 0 & 0 \\
    0 & 0 & 0
\end{pmatrix}
\]
where $Df(\xi^s)$ is the Jacobian matrix of the function $f$ estimated on the synchronization manifold $\xi^s$, $\gamma'(x^s)$ is the derivative of $\gamma(x)$ with respect to $x$ evaluated in $x^s$, the first component of the synchronization manifold, and $\alpha + \beta$ are the eigenvalues of the weighted connectivity matrix $G$.

The largest Lyapunov exponent of the MSE $\Lambda(\alpha, \beta)$ is known as master stability function and it permits to identify the synchronous conditions: a network synchronizes if and only if all the eigenvalues (apart the largest one) are in a region in which the MSE $\Lambda(\alpha, \beta)$ is negative.

In the study case the connectivity matrixes $C$ and $G$ are symmetric. It follows that the eigenvalues are real and the MSE, Eq. (9), and the MSF depend only on $\alpha$. In Eq. (9) another parameter can be identified: the product between $g_s$ and $k$. Defining $\eta = k g_s$, the MSE can be obtained as function of $\alpha$ and $\eta$, namely $\Lambda(\alpha, \eta)$, and the result is reported in Fig. 2.

In Fig. 3 the intersection between the MSE and the plane $\Lambda = 0$ is reported and the result is that the largest Lyapunov exponent is positive in the upper half-plane (with respect the $\Lambda = 0$ line) and negative otherwise. It follows that the studied networks of Hindmarsh-Rose neurons are Class-A networks (see [9]), namely the network synchronizes if $\lambda_2 < 0$ where $\alpha = a(\eta)$ is the $\Lambda = 0$ line and $\lambda_2$ is the second largest eigenvalue of the weighted connectivity matrix $G$, whose spectrum is composed of $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N$.

Note that in [4] qualitatively similar results have been obtained, but they are not numerically accurate as those reported in the following Section.

\[\text{Figure 2: MSF } \Lambda(\alpha, \eta) \text{ of the network of Hindmarsh-Rose neurons.}\]
\[\text{Figure 3: Sign of the MSF: Intersection between the MSF (reported in Fig. 2) and the plane } \Lambda = 0.\]
4. Synchronization Properties

Thanks to the results obtained in the previous Sec., fixed a network, i.e., its topology (described by $C$), its coupling strength $g_s$, and its coupling type, the synchronization conditions are known. In this Sec., the synchronization conditions of such a network are investigated in details.

Even if two different types of coupling are studied their MSEs/Fs turn out to be the same: the case of the electrical coupling can be obtained from the other by setting $k = 0$, i.e. $\eta = 0$. However, their properties are different:

- Electrical coupling ($k = 0$): the connectivity matrix is a Laplacian matrix and the number of connections of each cell can be arbitrary;
- Synaptic coupling ($k \neq 0$): the number of connections of each cell is equal to that of all the other cells.

In the first case (electrical coupling) the synchronous state is that of an isolated Hindmarsh-Rose cell, as it can be obtained from Eq. (8). Furthermore, the largest eigenvalue of the connectivity matrix $\lambda_1$, which is equal to 0 since the matrix is Laplacian, corresponds to the largest eigenvalue of the synchronous state, so only the other eigenvalues (which are all negative) have to be inside the region in which the MSF is negative. Therefore, the synchronization condition is
\[
\lambda_2 = g_s \lambda_2^* < a(0) = -0.68
\]  
(10)
where $\lambda_2^*$ is the second largest eigenvalue of the connectivity matrix, $\lambda_2$ is the second largest eigenvalue of the weighted one, and $a(\eta = 0)$ is obtained from the MSF reported in Fig. 3. Then, the following proposition is proved. 

**Proposition 1** For each network of Hindmarsh-Rose neurons defined as in Eq. (4) with electrical coupling there exists a minimum value of $g_s$, namely $g_s^* = a(0) / \lambda_2^*$ such that the network synchronizes for every $g_s > g_s^*$ and the synchronization behavior is equal to the one of an isolated Hindmarsh-Rose neuron.

In the case of synaptic coupling it is possible to see that the largest eigenvalue $\lambda_1^*$ of the connectivity matrix is equal to $k g_s$. Furthermore, this last eigenvalue corresponds to the largest Lyapunov exponent of the synchronous state. In fact, the MSE (9) is the variational equation of the synchronous state (8) in case of $\alpha = k g_s$ and $\beta = 0$. Then, only the other eigenvalues have to be inside the region in which the MSF is negative and the synchronization condition is
\[
\lambda_2 = g_s \lambda_2^* < a(\eta)
\]  
(11)
where $\lambda_2^*$ is the second largest eigenvalue of the connectivity matrix, that can be positive or negative, $\lambda_2$ is the second largest eigenvalue of the weighted one, and $a(\eta)$ is obtained from the MSF reported in Fig. 3. So, the following proposition is proved.

**Proposition 2** For each network of Hindmarsh-Rose neurons defined as in Eq. (4) with synaptic coupling, if $\lambda_2^*$ is negative (positive) there exists a minimum (maximum) value of $g_s$, namely $g_s^*$ such that the network synchronizes for every $g_s > g_s^*$ ($g_s < g_s^*$).

Note that in the case under examination (synaptic coupling) the synchronous equation (8) is not equal to the equation of an isolated Hindmarsh-Rose neuron (1). Then, if the synchronization condition (11) is verified the state variables of all the cells synchronize, but the synchronous behavior is not known. This fact motivates the investigation of the dynamic behavior, i.e. the bifurcation diagram with respect the parameter $\eta$, of the system reported in Eq. (8).

For lack of space, the bifurcation diagram is not reported but some of the obtained results are shown and explained. There is a critical value of $\eta$, namely $\eta^*$, such that if $\eta < \eta^*$ then the synchronous behavior is qualitatively equal to the one of the isolated Hindmarsh-Rose neuron, i.e. spikes during the burst phases. In case of $\eta > \eta^*$ the synchronous behavior is qualitatively different to the one of the isolated Hindmarsh-Rose neuron. For example it is possible to observe damped oscillations during the burst phases, burst phases without oscillations, or loss of burst phases.

A simple example follows where a network composed of only two Hindmarsh-Rose neurons is considered and the following connectivity matrix is used
\[
C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}
\]  
(12)
from which $k = 1$, $\lambda_1 = 1$, and $\lambda_2 = -1$. Fig. 4 shows the quadratic error between the corresponding state variables of the two cells and the time evolution of the state component $\chi_1(t)$ of the first cell for the coupling parameters $g_1^{(1)} = 1.24667$, $g_1^{(2)} = 1.2847$, and $g_2^{(1)} = 2.0$ respectively. In the following table all the parameters are reported and it is possible to note that in all three cases the network synchronizes, as shown in the above cited figures. Furthermore, the time evolution of the synchronous state only in the case 1 is composed of spikes in the burst phases, whereas in the other two cases it has a different behavior.

<table>
<thead>
<tr>
<th>Case</th>
<th>$g_s$</th>
<th>$\eta = k g_s$</th>
<th>$a(\eta)$</th>
<th>$\lambda_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.24667</td>
<td>1.24667</td>
<td>-0.93</td>
<td>-1.24667</td>
</tr>
<tr>
<td>2</td>
<td>1.2847</td>
<td>1.2847</td>
<td>+1.22</td>
<td>-1.2847</td>
</tr>
<tr>
<td>3</td>
<td>2.0</td>
<td>2.0</td>
<td>+2.00</td>
<td>-2.0</td>
</tr>
</tbody>
</table>

Thanks to these considerations Proposition 2 can be completed with the following lemma.

**Lemma 3** For the modified Hindmarsh-Rose neuron model defined as in Eq. (8) there exists a maximum value of $g_s$, namely $g_s^M$, such that the synchronization behavior is qualitatively equal to the one of an isolated Hindmarsh-Rose neuron if $g_s < g_s^M$. 

673
5. Conclusion

Conditions of synchronization in networks of Hindmarsh-Rose neurons have been obtained, as reported in Prop. 1 and 2, for two different types of coupling: electrical and synaptic. These conditions are more accurate than the results reported in other papers in literature. Furthermore, it has been shown that the bifurcation diagram of the synchronous state have an important role in the considered study, as remarked with Lemma 3.

6. Acknowledgement

This work was supported in part by Ministero dell’Istruzione, dell’Universit`e e della Ricerca under PRIN Project no. 2004092944.J04.

References


Queues, Long–range dependent traffic and non–linear maps

R. J. Mondragón † and D. K. Arrowsmith‡
†Dept. of Electronic Engineering, ‡School of Mathematical Sciences,
Queen Mary, University of London
Mile End Road, London E1 4NS, UK
Email: r.j.mondragon@elec.qmul.ac.uk, d.k.arrowsmith@qmul.ac.uk

Abstract—We present an overview of the progress made in the analysis of self-similar traffic models and, in particular, the queue performance using chaotic maps. Data packet traffic can be described by an ON/OFF (active/silent) model where ON means traffic production. If the ON periods are described by a heavy–tailed distribution then the traffic trace is long–range dependent. If this traffic is fed to a queue, the decay of the queue length probability is also heavy–tailed. In this paper we show that it is possible to change the queue length probability from heavy–tailed to exponential without destroying the long–range dependence property of the traffic.

1. Introduction

The queuing performance of traffic depends heavily on the statistical nature of traffic. It has long been known that internet traffic is long–range dependent (LRD) [6]. One definition of LRD is that a second–order stationary time–series has LRD if the sum of its autocorrelation function \( C(L) \) over all time lags \( L = 0,1, \ldots, \sum_{l=0}^{\infty} C(L) \), diverges. Often a specific asymptotic form is assumed for the fall off \( C(L) \stackrel{L \to \infty}{\sim} L^{2H-2} \), where \( L \to \infty \) means asymptotically equal to, up to a multiplicative constant, as \( L \to \infty \). The Hurst parameter \( H \in (1/2, 1) \) is the parameter which defines LRD. A higher value of \( H \) represents stronger correlations and \( H = 1/2 \) represents independent or short–range dependent data. Long range dependent traffic has several interesting properties that it can have significantly worse queuing performance than non–LRD traffic; when independent traffic streams are multiplexed, the resultant traffic has the highest Hurst parameter of all the streams multiplexed.

Attempts to improve LRD packet traffic as it crosses a network run up against an obstacle. The LRD properties of the traffic are robust against control techniques like traffic shaping, resource management, call admission control and priority control [9, 7, 2]. These network operations do not change the LRD properties and therefore it might be assumed that their effect on the QoS is limited. However, in this article, we present an example to show that this is not an insurmountable obstacle.

Packet traffic can be modelled using an ON/OFF (or active/silent) model also known as a packet train model [5] (see fig. 1). The ON state relates to the period of continuous traffic activity. If the probabilities of the duration of the active region silent region

\[
\begin{array}{c}
\text{packet train} \\
011111000010
\end{array}
\]

Figure 1: A packet train and its representation as a binary sequence.

ON or/and OFF decay as a power law (heavy–tailed distribution) then the auto–correlation function of the ON/OFF will also decay slowly, making the traffic LRD [4]. If the traffic arriving to the queue has a heavy–tailed ON time then the queue length probability will be heavy–tailed [11]. In this paper we show that this is not the only way of producing LRD traffic, it is perfectly possible to have LRD with no power laws in the ON/OFF periods. Hence, it is possible to construct or modify an ON/OFF traffic trace such that the periods of continuous activity/inactivity are short but the traffic is LRD, and more remarkably, a queue fed by this traffic will have exponentially decaying queue lengths. The importance of this result is that it shows that LRD is not a straightjacket for the traffic control mechanisms.

2. Chaotic Maps as Models of Packet Traffic

The traffic is modelled with an ON/OFF source described by the family of intermittency maps [3]

\[
x_{n+1} = F(x_n; d, m_1, m_2)
\]

\[
\begin{align*}
F_1(x_n) & = x_n + ((1 - d)/d^m) x_n^m, \quad 0 < x_n < d \\
F_2(x_n) & = x_n - (d/(1 - d)) x_n^{m_2}(1 - x_n)^{m_2}, \quad d < x_n < 1,
\end{align*}
\]

(1)

where \( x_n \in (0,1) \) and parameters \( m_1, m_2 \in (1,2) \) and \( d \in I=(0,1) \). The real sequence \( \{x_n\} \) can easily be turned into a binary sequence with the use of an output map \( y: I \to \{0,1\} \) defined by

\[
y_n = y(x_n) = \begin{cases} 
0 & 0 < x_n \leq d, \quad \text{OFF (silent region)} \\
1 & d < x_n < 1, \quad \text{ON (active region)}. 
\end{cases}
\]

(2)
The binary sequence represents the creation of traffic. As the duration of the active/silent regions can be manipulated individually the maps can produce mono-fractal traffic (like fractional Brownian models) and multi-fractal traffic [8]. If \( m_1, m_2 > 3/2 \) the traffic generated by eqs. (1, 2) is an asymptotically second order self-similar process. If \( m_1 = m_2 = 1 \) the maps generate a Poisson process with a geometrical decay in its active and silent region and the traffic is not correlated.

If \( \sigma_A \) denotes the number of iterations that an orbit spends in the active region then, “heavy tailed” behaviour means that

\[
\mathcal{A}(K) = \text{Prob}(\sigma_A = K) \sim K^{-m_1/(m_1-1)}, m_2 > 3/2
\]  

[3, 10, 8]. Similarly if \( \sigma_S \) is the number of iterations that an orbit spends in the silent region then

\[
\mathcal{S}(K) = \text{Prob}(\sigma_S = K) \sim K^{-m_1/(m_1-1)}, m_1 > 3/2.
\]  

If \( 3/2 < m_1, m_2 < 2 \) and \( m = \max(m_1, m_2) \) then, the auto-correlation is \( C(L) \sim L^{-m_1(2m - 2)/(m-1)} \), where \( L \) is the lag [10, 8, 1].

3. Chaotic Maps and Queues

The queue length dynamics \( Q_n \) at time \( n \) of a deterministic server is described by

\[
Q_{n+1} = Q_n - \Theta(Q_n) + M_{n+1}
\]  

where \( M_{n+1} \) is the number of queue arrivals at time \( n + 1 \), \( \Theta(Q_n) \) is the Heaviside function \( \Theta(Q_n > 0) = 1 \) and \( \Theta(Q_n = 0) = 0 \) and it describes the queue departures. The arrivals are described by a chaotic map of the form [3]

\[
M_{n+1} = \mathcal{R}(M_n) = N_{ik} y(x_n),
\]  

where \( y(x_n) \) is the indicator variable given by Eqs. (1,2) and \( N_k \) is the number of packet arrivals at a given time.

The dynamics of the arrivals is given by eq. (1) hence the evolution of the traffic source and queue length is given by a deterministic transformation in the variables \( (x_n, q_n) \) (see fig. 2). The dynamics of the system is captured in the “surface of section” \( x = d \). The dynamics of the queue is split into silent periods, where there are no arrivals, and active periods. The duration of the \( n \)th silent period \( \sigma_{q,n}^{s} \), is the time that an orbit of the map \( F \) spends in \( x \in (0, d) \). Similarly, the duration of the \( n \)th active period \( \sigma_{q,n}^{a} \), is the time that an orbit spends in \( x \in (d, 1) \) where the map generates traffic. When inside the active period at each iteration of the map \( \mathcal{R} \), there is \( N_i \geq 1 \) packets arriving to the queue. The total number of packets that arrive at the queue in the active period is \( \sum_{i=1}^{\sigma_{q,n}^{a}} N_i \).

The up \( U_n \) and down \( D_n \) map describe the dynamics just before crossing the surface of section (see fig. 1). The up-map gives the queue length just before the silent period is going to start. The down-map is the length of the queue at the end of an active period minus the number of packets served in the silent period. If the iterate of the down–map is non–negative then the queue length is given by the down map. If the iterate of the down–map is negative, then the queue length is zero. The dynamics of the up and down map is

\[
U_{n+1} = \left( \sum_{i=1}^{\sigma_{q,n}^{a}} N_i - \sigma_{q,n}^{a} \right) + D_n \Theta(D_n)
\]  

\[
D_{n+1} = U_{n+1} - \sigma_{q,n}^{a}.
\]  

If \( \mathcal{P}(D \Theta(D) = l) \) is the probability that at the end of a silent period the queue length is \( l \) and \( X_m = \mathcal{P}(\sum_{i=1}^{\sigma_{q,n}^{a}} N_i - \sigma_{q,n}^{a} = m) \) is the probability of having an excess of \( m \) packets at the end of the \( n \)th active period then

\[
\mathcal{P}(U = K) = \sum_{l=0}^{K} \mathcal{P} \left( \sum_{i=1}^{\sigma_{q,n}^{a}} N_i - \sigma_{q,n}^{a} = K - l \right) \mathcal{P}(D \Theta(D) = l).
\]  

Using the properties of the step function, eq.(7) and the notation \( \mathcal{U}_K \equiv \mathcal{P}(U = K) \) then

\[
\mathcal{U}_K = X_K \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} S_j \mathcal{U}_l + \sum_{i=0}^{K-1} X_{K-i} \sum_{j=0}^{\infty} \sum_{l=0}^{\infty} S_j \mathcal{U}_l S_j,
\]  

gives the probability that the value of the up map is exactly \( K \).

3.1. Case \( N_n = 2 \) for all \( n \)

In this case there are two packets arrivals to the queue per iterate when inside the active region. The total number of packets generated in the active region is \( 2\sigma_A \) where \( \sigma_A \) is the duration of the active region. Then

\[
X_K = \text{Prob} \left( \sum_{i=1}^{\sigma_{q,n}^{a}} N_i - \sigma_{q,n}^{a} = K \right) = \text{Prob}(2\sigma_A - \sigma_A = K) = \text{Prob}(\sigma_A = K) = \mathcal{A}_K
\]  

Figure 2: Orbit of the Queue-Intermittency map dynamics and the relationship between the queue dynamics and the up \( U_n \) and down \( D_n \) map.
\[ U_K = \mathcal{A}_K \sum_{i=0}^{\infty} U_i \sum_{j=1}^{K-i} \mathcal{A}_{K-i} \sum_{j=0}^{\infty} U_{i+j} S_j, \tag{11} \]

If the active region decays as \( \mathcal{A}_K \xrightarrow{K \to \infty} \kappa^\alpha \) then \( U_K \xrightarrow{K \to \infty} \kappa^\alpha \delta^K \) regardless of whether the silent region decays as a power law or an exponential. The case where \( \mathcal{A}_K \xrightarrow{K \to \infty} \kappa^\delta \) implies \( U_K \xrightarrow{K \to \infty} \kappa^\delta \delta^K \) \cite{[8]}. For the case that the probabilities of the active and silent region decay as an exponential, queue probability decays exponentially fast. The queue length probability decays as a power law as a consequence of heavy–tailed ON times. If the probability that the duration of the ON time is changed from power–law to exponential decay then, we can expect that the queue length probability will change from a power law to an exponential decay.

4. The Hiccup Map

It is possible to construct a packet traffic model that will produce traffic traces with both ON and OFF times decaying exponentially fast and with an auto–correlation function that decays as a power law. The iterates \( b_n \) of the Bernoulli map \( B \)

\[ b_{n+1} = B(b_n) = \begin{cases} b_n, & 0 < b_n < d_b \\ d_b - b_n, & d_b < b_n < 1 \end{cases} \tag{12} \]

are used to decide if the intermittency map is iterated or not,

\[ x_{n+1} = H(x_n) = \begin{cases} F_1(x_n), & 0 < x_n < d \\ x_n, & d < x_n < 1, 0 < b_n < d_b, \\ F_2(x_n), & d < x_n < 1, d_b < b_n < 1. \end{cases} \tag{13} \]

The value of \( x_{n+1} = H(x_n) \) remains constant at \( x_n \) if \( 0 < b_n < d_b \). We interpret this case as the event where no packets are created and a gap in the ON train is introduced. The average number of spaces that are introduced in the ON region is determined by the average number of iterates that an orbit of the Bernoulli map spends in the \( 0 < x < d_b \) region which simply is \( \eta = (1 - d_b) \).

The sequence of symbols that represents the traffic has to be modified to include the creation of spaces in the ON period arising from the Bernoulli orbits. To distinguish these spaces with the OFF period events we denote the injected spaces by the symbol \( \bar{\theta} \). The indicator map becomes

\[ y_{n+1} = \begin{cases} 0, & 0 < x_{n+1} < d \\ \bar{\theta}, & d < x_{n+1} < 1; d < x_n < 1; 0 < b_n < d_b \\ 1, & d < x_{n+1} < 1; d < x_n < 1; d_b < b_n < 1 \end{cases} \tag{14} \]

where the silent region is given by \( 0 < x_n < d \) and an extended active region by \( d < x_n < 1 \). Notice that in the sequence of symbols a \( \bar{\theta} \) and a 0 never appear consecutively.

The hiccup map breaks the long packet trains into a “convoy” of smaller trains, with the property that the convoy travels together, that is they are correlated.

5. The Queue and the Hiccup Map

As in the case of the intermittency map, it is possible to capture the Hiccup–queue dynamics by studying the behaviour of the queue when the intermittency map changes from the active to the silent region. However, the active region of the Hiccup map sometimes does not generates traffic so there is a possibility that the queue gets emptied. To consider this possibility the up and down map are modified such that

\[ U_{n+1} = \left( \sum_{i=1}^{2\theta} N_i - \tau_{A_n} \right) \Theta \left( \sum_{i=1}^{2\theta} N_i - \tau_{A_n} \right) + D_n \Theta(D_n) \]

\[ D_{n+1} = U_{n+1} - \sigma_{S_n}, \tag{15} \]

where the term \( \Theta(\sum_{i=1}^{2\theta} N_i - \tau_{A_n}) \) describes the probability of having an empty queue at the end of the extended active region \( A_n \).

5.0.1. Case \( N_i = 2 \) for all \( n \)

The probability of the up–map is

\[ \text{Prob}(U = K) = \sum_{i=0}^{K} \text{Prob}((2\sigma_A - \tau_A)\Theta(2\sigma_A - \tau_A) = K - i) \text{Prob}(D\Theta(D) = i) \]

\[ \text{Prob}(D = K) = \sum_{i=0}^{K} \text{Prob}(2\sigma_A - \tau_A = i) \text{Prob}(\Theta(D) = i) \]

where \( \sigma_A \) is the duration of the active region of the intermittency map and \( \tau_A \) is the duration of the extended active region for the Hiccup map. Notice that \( \tau_A \geq \sigma_A \). If \( X_m \equiv \text{Prob}(2\sigma_A - \tau_A = m) \) is the probability of having an excess of \( m \) packets at the end of the \( n \)th extended period, then from eq. (16)

\[ U_K = X_K \sum_{i=0}^{\infty} U_i \sum_{j=0}^{\infty} S_j + \sum_{i=1}^{K} X_{K-i} \sum_{j=0}^{\infty} U_{i+j} S_j \]

\[ + \sum_{i=1}^{\infty} X_{i-1} \sum_{j=0}^{\infty} U_{K+i+j} S_k. \tag{17} \]

For the intermittency–queue maps, the behaviour of the \( \mathcal{U}_K \) as \( K \to \infty \) is related to the decay of \( X = \mathcal{X}_K \) (see eq. (11)). In the Hiccup–queue map, we have an equivalent case, the decay of the queue is related to the decay of \( X_K \). To evaluate \( X_K \), it should be noted that the probability of
the extended active region lasting $\tau_A$ iterates if the original active region is $\sigma_A$ is $\eta\sigma_A(1-\eta)\tau_A-\sigma_A$.

To evaluate if at the end of an active region there is a surplus or deficit of packets in the queue we need to evaluate that, given a time $\sigma_A$, what is the probability that the extended region lasts $\tau_A$ when the surplus/deficit of packets is $m = 2\sigma_A - \tau_A$ hence

$$\mathcal{A}_L = \eta \sum_{i=0}^{L-2} \mathcal{A}_{L-i} \eta^{i} \left( L - \frac{2}{i} \right)$$

The probability of having an excess of $L$ packets at the end of the active region is

$$X_L = \sum_{i=L}^{\infty} \mathcal{A}_{i} \eta^{i-1} \left( 1-\eta \right)^{i-L} \left( 2i - L - 2 \right) + \mathcal{A}_L \delta_{i,L}$$

where $\delta_{i,L}$ if $L = 1$. Notice that if $\mathcal{A}_{i} \sim \eta^{-\gamma}$ then the first sum can be bounded using $X_L < \mathcal{A}_L \sum_{i=L}^{\infty} \eta^{-i} \left( 1-\eta \right)^{i-L} \left( 2i - L - 2 \right)$. The summation part of the bound can be simplified to

$$\sum_{i=L}^{\infty} \eta^{-i} \left( 1-\eta \right)^{i-L} \left( 2i - L - 2 \right) = \eta^{L-1} H \left( \frac{L}{2} \times \frac{1}{2} \times \frac{1}{2} ; L-1 ; 4 \eta \left( 1-\eta \right) \right)$$

The Hypergeometric function $H(\cdot ; \cdot ; \cdot)$ has a singular point at $z = 4 \eta (1-\eta) = 1$ or $\eta = 1/2$. Expanding the previous equation for $\eta = 1/2 + \epsilon$, with $\epsilon > 0$ then

$$X_L \overset{L \to \infty}{\sim} \mathcal{A}_L \left( \frac{1}{2} + \frac{1}{4\epsilon} \right)$$

which means the queue decays as a power law (given by the decay of $\mathcal{A}_L$). For the case $\eta = 1/2 - \epsilon$, with $\epsilon > 0$

$$X_L \overset{L \to \infty}{\sim} \mathcal{A}_L \left( \frac{e^{-(1/2c)} \sinh(1/2c)}{2c(1 - 2c)} \right)$$

which means the queue decays exponentially fast, independently of $\mathcal{A}_L$. This is the main result of this paper (see fig. 3).

For a simple queuing system we found an analytical expression that shows the power law decay of the queue can be abruptly changed to an exponential decay by modifying the properties of the ON times of the traffic without destroying its LRD characteristics. This modification is achieved by a random interruption of the iteration of the active region, where an interruption also means the stoppage of creation of data. We believe that this result can be useful in the control of packet traffic networks.

Acknowledgements

This work was partially supported by the Engineering and Physical Sciences Research Council, UK (EP/C520246).

References


An autonomous Piecewise-Constant System with Multirate Sample-Hold Controllers

Tadashi TSUBONE and Yasuhiro WADA

Department of Electrical Engineering, Nagaoka University of Technology
1603-1 Kamitomioka-cho, Nagaoka, Niigata, 940-2188, Japan
Phone:+81-258-47-9558, Fax:+81-258-47-9500,
Email: tsubone@vos.nagaokaut.ac.jp

1. Introduction

Synthesis of chaotic oscillator with simple structure have received grate attention due to the increasing number of application of chaotic oscillator in engineering, as for example in communication systems [1].

Several models of simple chaotic oscillator with switching elements have been reported in some literature because of their simple structure allowing the implementation without difficulty and the use of less complex tools for its analysis.

On the other hand, periodic solutions are one of the most important phenomena in nonlinear dynamical system including many engineering systems. While stability analysis of periodic solutions is a basic problem, synthesizing a nonlinear system which exhibits a stable periodic solution is also essential. Several methods for the inverse problem of synthesizing such systems have been proposed [2]-[3].

In this paper, we propose a novel nonlinear system which consists of a chaotic system and multirate sample-hold controllers. The proposed system exhibits some stabilized Unstable Periodic Orbits (abbr. UPO) which are embedded on chaos attractor of the original chaotic system. Generally, the procedure to stabilize UPOs is called Controlling Chaos. Our proposal is a synthesis of a nonlinear system which generates periodic solutions based on chaos.

The basic principle is using a feedback of delayed states, but is not included on a category of Delayed Feedback Control (abbr. DFC) [4]-[5] well-known as one of a controlling chaos methods. However a proposed system has an advantage such that no preliminary calculation of the UPOs is required, similarly to the DFC.

First, we consider an autonomous piecewise-constant system [6]-[7] with sample-hold controllers as the basic chaotic system. The system dynamics is governed by a 1-D return map. By using the return map, we can accomplish the both of synthesis and analysis of the system. Chaos generation can be guaranteed theoretically. Second, we construct a controlled system based on the return map. We provide a condition to stabilize UPOs and a domain of attraction. Some theoretical results are verified in the experimental circuit.

2. An autonomous piecewise-constant system with a sample-hold controller

An autonomous piecewise-constant system with a sample-hold controller is shown in Fig. 1, where the sampling switch samples $x$ at any zero-hitting moment of $y$ and the holder $H$ stores the sampled value until the next zero-hitting of $y$ occurs. Let the $n$-th zero-hitting moment of $y$ be $\tau_n$ and let the output of $H$ be $x_s$:

$$x_s = x(\tau_n), \quad \text{for} \quad \tau_n \leq \tau < \tau_{n+1}, \quad (n = 0, 1, 2, 3, \cdots).$$

(1)

Then the system dynamics is described as the following:

$$\dot{x} = \text{sgn}(y)$$
$$\dot{y} = -\text{sgn}(x-h)$$
$$h = x_s + a(x_s - \text{sgn}(x_s)).$$

(2)

where $\cdot$ denotes differentiation by normalized time $\tau$ and the nonlinear function $\text{sgn}(\cdot)$ is defined as the following:

$$\text{sgn}(X) = \begin{cases} 
1 & \text{for } X \geq 0, \\
-1 & \text{for } X < 0.
\end{cases}$$

(3)

Since the time-invariant function $\text{sgn}$ takes only constant value, the right hand side of the equation is piecewise-
constant and the trajectory of the solution is piecewise-linear. Figure 2 is an example of the trajectory in a phase space. Here, let the n-th x be xn, namely xn is x(τn).

The trajectory started from xn moves away from x-axis and heads to a threshold x = h0. Then it must reach a point (h0, h0 - x0) on the threshold, where

\[ h_n = x_n + a[x_n - \text{sgn}(x_n)]. \]

And the trajectory leaved from the threshold must return to x-axis. Then it must reach a point (2h0 - x0, 0) on x-axis. Here, focusing the state xn at \( \tau = \tau_n \), we can define the return map \( F(x_n) \) from xn to x_{n+1}. The return map can be described explicitly as the following:

\[ F : x_n \mapsto x_{n+1}, \quad F(x_n) = 2h_0 - x_n. \]

Changing the nonlinear function to determine the threshold \( h_0 \), we can construct any 1 dimensional (abbr. 1-D) return map \( F(x) \). In this paper, we consider the case that \( h_0 \) is defined by (4).

We assume that the unique parameter a is constant and \( a = 2 \) after here. In this case, Substituting (4) and \( a = 2 \) into (5), the return map \( F(x_n) \) can be described and is to be a well-known Bernoulli shift map as the following:

\[ F(x_n) = \begin{cases} 2(x_n - 1) + 1, & \text{for } x_n \geq 0, \\ 2(x_n + 1) - 1, & \text{for } x_n < 0. \end{cases} \]

In the case of \( a = 2 \), the discrete time system (6) exhibits chaos. In the experimental circuit, the integrator in the Fig. 1 is realized by using R, C and OP-amp. (TL074), and S/H is composed of a capacitor and analog switches (4066). The output voltage v of the integrator and \( v_n \) of S/H corresponds to x and \( x_n \), respectively. Figure 3 shows the chaos attractor on the phase space and the corresponding laboratory measurement. Figure 4 shows the chaos attractor of the return map (6).

3. A System with Multirate Sample-Hold Controllers

We introduce some basic definitions for the return map.

**Definition:** A point \( x_i \) is said to be a l-periodic point if \( F^l(x_i) = x_i \) and \( F^k(x_i) \neq x_i \) for 0 < k < l, where \( F^l \) denotes the l-fold composition of \( F \). A l-periodic point \( x_i \) is said to be stable (respectively, unstable) if \( |DF^l(x_i)| < 1 \) (respectively, \( |DF^l(x_i)| > 1 \)), where \( DF^l \) denotes derivative of \( F^l \).

A sequence of a l-periodic point, \( \{x_1, F(x_1), \cdots, F^{l-1}(x_1)\} \), is said to be a periodic sequence with period l. We refer to a stable periodic sequence as a periodic attractor. Hereafter we abbreviate an unstable periodic point by UPP and abbreviate an unstable periodic sequence with l-period by l-p UPS.

Figure 5(a) shows a block diagram of a system with Multirate Sample-Hold Controllers and the timing charts for the switches in this diagram. The switches are closed if the clock signal shown as the charts in Fig. 5(b), is high level. The pulse signal \( CK_1 \) turns to high level at the moment of \( y = 0 \), and the \( CK_2 \) turns to high level at every \( l \)-th pulse of \( CK_2 \). Also, the switch controlled by \( S_{\text{CTL}} \) is governed by the following manner

\[ S_{\text{CTL}} : \begin{cases} \text{ON}, & \text{for } \tau_m \leq \tau < \tau_{m+1}, \\ \text{OFF}, & \text{for } \tau_{m+1} \leq \tau < \tau_{(m+1)}. \end{cases} \]

The dynamic of the System with Multirate Sample-Hold...
Controllers is described by follows:

\[
\begin{align*}
\dot{x} &= \text{sgn}(y) \\
\dot{y} &= -\text{sgn}(x-h) \\
h &= \frac{1}{\tau}(x_1 + z_1) + a(z_3 - \text{sgn}(z_3)), \quad \text{for } ln \leq \tau < ln + 1, \\
\dot{x} &= \text{sgn}(y) \\
\dot{y} &= -\text{sgn}(x-h) \\
h &= x_3 + a(x_1 - \text{sgn}(x_1)), \quad \text{for } ln + 1 \leq \tau < l(n + 1).
\end{align*}
\]

where \( z_1 = (1 - K)x_3 + K \cdot y_r \), \( y_r \) is an output of the holder of the controller and \( K \) is a control gain. By using similar procedure in Sec. 2, we can derive the corresponding discreet-time system as the following:

\[
x_{n+1} = \begin{cases} 
F((1 - K)x_0 + K y_{n-1}), & \text{for } n = kl, \\
F(x_0), & \text{for otherwise}
\end{cases}
\]

\[
k = 1, 2, \ldots
\]

The return map from \((x(ln), y(ln))\) at \( \tau = ln \) to \((x(l(n + 1)), y(l(n + 1)))\) at \( \tau = l(n + 1) \) can be given by

\[
x_{n+1} = F^l((1 - K)x_n + K \cdot y_n) \\
y_{n+1} = (1 - K)x_n + K \cdot y_n.
\]

Note that if \( x_n \) is identical to \( y_n \), the discreet-time system (10) is equivalent to (5).

Let \( \xi \) be one UPP of \( l\)-p UPS of the system (5). letting \( \hat{x}_n \) be defined by \( x_n - \xi \), letting \( \hat{y}_n \) be defined by \( y_n - \xi \) and letting \( A \equiv (\delta / \delta x_0)F^1_{l\xi-n\xi} \), the linearized system of (10) in the neighbor of \( \xi \) is described by

\[
\begin{bmatrix}
\dot{x}(n + l) \\
\dot{y}(n + l)
\end{bmatrix} =
\begin{bmatrix}
A(1 - K) & AK \\
1 - K & K
\end{bmatrix}
\begin{bmatrix}
x(n) \\
y(n)
\end{bmatrix}.
\]

Here, if we set the gain of a controller to

\[K = -A(1 - A)^{-1},\]

then all of the characteristic root of the linearized system (11) are identical to zero, that is, the system is to be stable in the neighbor of \( \xi \). The solution \( x_n \) started from the neighbor of \( \xi \) at \( \tau = 0 \) must converge to \( l\)-p UPP.

Figure 6 shows a simulation result of generating 3 periodic solutions by setting parameters as \( l = 3 \) and \( K = \frac{8}{7} \). Figure 6(a) and 6(b) shows stabilized 3-p UPPs which have different initial conditions. 1 periodic solutions are also observed on same parameters. The solution is identical to the unstable fixed point of the original chaotic system (5) \( x_\infty = -1 \) and \( x_\infty = 1 \).

4. The domain of attraction

First, in order to classify the periodic solution, we define the binary sequence corresponding to a UPS \( P(x_0) = \{x_0, F(x_0), \ldots, F^{l-1}(x_0)\} \) as the following:

\[
S(x_0) = \{\omega_0(x_0), \omega_1(x_0), \ldots, \omega^{l-1}(x_0)\},
\]

\[
\omega^m(x) = \begin{cases} 0, & \text{if } 0 \leq F^m(x) < \frac{1}{2} \\
1, & \text{if } \frac{1}{2} \leq F^m(x) < 1
\end{cases}
\]

\[
681
\]
And let the binary sequence
\[ S(x_l) = \{00\ldots011\ldots111\ldots1\} \]
be represented by \((0^n 1^i \ldots 1^s)\). If the \(S(x_l)\) corresponding to \(l\)-p UPS \(P(x_l)\) is \((0^n 1^i \ldots 1^s)\), we define that \(P(x_l)\) is \((0^n 1^i \ldots 1^s)\) UPS. An attractor shown in Fig. 6(a) is the case of \((011)1\) UPS. Similarly, an attractor shown in Fig. 6(b), 6(c) and 6(d) is \((01)1\) UPS, \((01)3\) UPS and \((1)1\) UPS, respectively. Note that \((11)1\) UPS is same to \((1^2)0\) and \((101)1\) UPS. Figure 7 shows the domain of attraction for some UPSs with \(l = 3\) which the solution started from an initial state \((x_0, y_0)\) converges to some periodic attractors. Figure 7(a), (b) and (c) is the case of \(K = \frac{1}{3}\), \(K = \frac{1}{3} + 0.05\) and \(K = \frac{1}{3} + 0.1\), respectively. The domain of attraction can be obtained theoretically and the detail will be discussed elsewhere.

5. Conclusions

We proposed a novel nonlinear system which consists of a chaotic system and multirate sample-hold controllers. The proposed system exhibits some stabilized Unstable Periodic Orbits (abbr. UPO) which are embedded on chaos attractor of the original chaotic system. Now we try the generalization of the system and consider engineering applications.

References


Fractal Structures of Three-Segment Piecewise-Linear Map in Reverse Time

A.I. Tomashevskiy†

† Department of Generation of Oscillations and Signals, Moscow Power Engineering Institute (Technical University),
14 Krasnokazarmennaya str., Moscow, Russia
Email: alex_tomash@mail.ru

Abstract – The behavior in reverse time of 1-dimension noninvertible discrete dynamical system in the form of 3-segment piecewise-linear map, demonstrating chaotic dynamics in direct time, is considered. The features of fractal distribution in phase space of reverse iteration points are investigated. Methods of formation of sets, having the structures of geometrical fractals, by means of special ordering are proposed.

1. Introduction

When investigating the nonlinear dynamical systems, the traditional approach is usually applied, representing the analysis of evolution in dynamics: the state of the system in given moment of time, i.e. initial condition, uniquely defines its state in subsequent moment. In other words, “direct problem of nonlinear dynamics” is solved [1]. Except for direct problem, there is also “inverse problem of nonlinear dynamics”. In classical form it consists of the searching of the state of the system in the moment, which is previous with respect to present moment, i.e. the determination of initial condition of the direct problem [2]. The inverse problem of chaotic dynamics has the exceeding difficulty of research because of many-valued solution. A noninvertible map, which demonstrates chaos in direct time, in reverse time may generate a set of system states, corresponding to the same moment of time, and their number increases with reverse iterations.

In this case, the solution of inverse problem is impossible without any additional information. However, attempts of examination of structures of reverse iteration points manifolds reduce to unexpected results. The structure of many-valuedness of noninvertible dynamical systems in reverse discrete time have been researched in previous works. As examples, we considered 1-dimension 1-parameter skew tent (two-segment piecewise-linear) [3] and skew logistic (two-segment piecewise-quadratic) [4] maps. It has been shown that processes in reverse time in these systems under certain conditions has complicated space-time structure: in the time domain it is chaotic behavior, but in space domain it has fractal distribution in the basin of attraction of corresponding chaotic attractor in direct time. In [4] we also proposed the convenient approach to revealing the fractal structures in reverse time, which consists of the ordering methods of reverse iteration points, resulting in the pronounced geometrical fractals.

In above-mentioned papers inverse discrete maps with two-segment nonlinearity, i.e. with two-valuedness, have been examined. In the present paper the analogous results, obtained for three-segment piecewise-linear map [5], is proposed for consideration. The extension of ordering principle to three-valued systems in reverse time is also presented. The attention is also paid to interesting phenomenon of formation in the system well-known fractal curve “Devil’s staircase” [5,6].

2. Three-segment piecewise-linear map in direct time

The system function \( F(x) \) of discrete map \( x_{k+1} = F(x_k) \), \( k=0,1,...,N \), is formed by 3 different pieces of lines joint in two points (A and C), Fig. 1,a. It can be written in the following way

\[
F(x) = \begin{cases} 
\frac{b}{x}, & 0 \leq x \leq a, \\
\frac{b(c-x)+d(x-a)}{c-a}, & a < x \leq c, \\
\frac{x-c+d(1-x)}{1-c}, & c < x \leq 1.
\end{cases}
\]

Function \( F(x) \) is always fixed in two points \((0,0), (1,1)\) and has four different parameters: the 1\(^{st}\) pair \((a,b)\) defines the position of A, the 2\(^{nd}\) \((c,d)\) – position of C; \( F(a)=b, F(c)=d, c\geq a, a,b,c,d \in [0,1] \). The definitional domain is given by \( x \in [0,1] \). The number of fixed points is equaled to either two or three: \( x_1^* = 0, x_2^* = 1, x_3^* = \frac{(ad-cb)}{(a+d-c-b)} \); the latter exists only if the middle segment of \( F(x) \) crosses the bisectrix inside the interval \((a,c)\), as shown Fig. 1,a. The analogous system in direct time have been investigated in [5].

If \( b>d \), then system function will noninvertible and will have two extremums: A is a maximum, C is a minimum. If at the same time \( x_3^* \) exists and its multiplier \( |\mu_3| = (b-d)/(c-a) > 1 \), then all fixed points will become unstable and dynamics of \((1)\) is chaotic practically for any parameters and initial conditions except for some special cases. In Fig. 1,b,c an example of chaotic process and its density of distribution are shown. The size of attractor is defined by \( b,d \) and for given case is \([0.1, 0.95]\), its basin of attraction coincide with the whole definitional domain; Lyapunov exponent of the attractor \( \Lambda \equiv 0.8 \).
3. Processes in the system in reverse time

The inverse map \( x_{k}=F^{-1}(x_{k+1}) \) can be obtained from the direct one by inverting the system function. For the considered nonlinearity (1) this results in three-valued dependence (2). Therefore, for any \( x_{k+1} \) we have from 1 to 3 values of \( x_{k} \) depending on the \( x_{k+1} \) and system parameters in contrast to two-valued systems [3,4], which gives from 0 to 2 values of \( x_{k+1} \). For the given \( x_{k} \) on every step \( n \in [0,N] \) backwards in time the number of generative values \( x_{k,n} \) continuously increases. It’s illustrated graphically as the inverse Lamerey-Königs diagram in Fig. 2,a and “tree of reverse iterations” in Fig. 2,c.

The number of reverse iteration points increases with \( n \) and fill up the basin of attraction. The density of distribution \( P(x) \) of the set of points \( \{x_{k-N}\} \) is shown in Fig 2,b. Comparing it with the density of distribution of the corresponding direct process in Fig. 1,b, we make the conclusion that reverse iterations and direct chaotic dynamics generates statistically different processes.

When detail analyzing of the profiles of functions \( P(x) \) for different parameters their self-similarity is revealed, which tells us that a set \( \{x_{k-N}\} \) has a fractal structure. However, it is not evident after visual observation of \( P(x) \)-functions. For more careful investigation of fractal structures in this system we apply the ordering principle that was developed in [3,4] concerning two-valued maps in reverse time. Here we propose its modification regarding to noninvertible systems with three-valuedness in reverse time.

The principle represents the following. At first we calculate the symbolic dynamics of every branch of tree of reverse iterations (separate trajectory in Fig. 2,c) defined by

\[
F^{-1}(x) = \begin{cases} \frac{a}{b}x, & 0 \leq x \leq b, \\ \frac{cb-ad-(c-a)x}{b-d}, & d \leq x < b, \\ \frac{c-d+(1-c)x}{1-d}, & d < x \leq 1. \end{cases}
\]

When \( n=N \) we have a set of reverse iteration points (“top of the tree”) defined by

\[
x_{k-N,N}=F^{-n,N}(x_{N}),
\]

where \( F_{j+1,N}(x) \) is \((j+1)\)-th segment of \( n \)-order inverse repeat map function, \( j \in [0,3^{n}-1] \) is the branch index.

The principle represents the following. At first we calculate the symbolic dynamics of every branch of tree of reverse iterations (separate trajectory in Fig. 2,c)

\[
\{\text{TER}\}_1 = \{\phi_{1-N,N-1}\}_{\text{TER}},
\]

where each symbol of which is equaled to either 0, 1 or 2 – in a sense of ternary system of numeration, in contrast to two-valuedness, where it is binary. A symbolic dynamics uniquely connected with the value of \( x_{k-N,N} \) in the top of tree. Then we convert \( \{\text{TER}\}_1 \) into decimal form by the rule

\[
j = \sum_{n=1}^{\infty} j_{k-N,n} G(n),
\]
where $G(n)$ is ordering function, which can be chosen arbitrary. The calculated value of branch index $j$ is the significant ordering parameter of the method. Some symbolic dynamics cannot be realized in considered system. Therefore corresponding index $j$ and value of $x_{k,j}$ don’t exist. But for $b=1$ and $d=0$ in (2) any symbolic dynamics is reachable.

In Fig. 3 the results of application of the ordering method, called by us TDD (“Ternary to Decimal – along Direct time”), where $G(n)=3^{n-1}$, are shown for several values of parameters, $\bar{j}=j/(3^n-1)$ – normalized branch index ($0\leq\bar{j}\leq1$). They are “top tree structures” with pronounced features of geometrical fractals (below the figures values of fractal dimensions $d_F$ are given, calculated by using the covering method [6]).

Fig. 3,a consists of the ordered set of points corresponded to the points of the right border of tree of reverse iterations and the density of distribution in Fig. 2. A small fragment is similar to whole image.

The structure in Fig.3,b is called by us “Cross fractal”, it’s obtained for symmetrical map function. The fractal is also symmetrical and self-similar. This set as well as previous one consists of the separate points nowhere adjoined to each other, which number tends to infinity when $n\to\infty$. Therefore, the topological dimension of the set $d_T=0$, while the fractal dimension $d_F=1$.

The structure in Fig. 3,c represents the well-known fractal curve “Devil’s staircase” [5,6], it consists of “stairs”, which number tends to infinity, but the height – to zero, when scaling. It looks like the increasing on average function, but having zero derivative in every point; its fractal and topological dimensions are equalled: $d_T=d_F=1$.

The solid fractal curves with such features occur in our system always when $a=c$ and $b=1$, $d=0$. They always have a plane interval in the middle: $\bar{j} \in (1/3, 2/3)$. In the case of $a=c<0.5$ the structure is symmetrical, so as left ($\bar{j} \in (0, 1/3)$) and right ($\bar{j} \in (2/3, 1)$) parts are identical. For other cases the curve becomes asymmetrical, if $a=c<0.5$, then values of $x_{k,j}$ will decrease, if $a=c>0.5$ they will increase.

The family of all possible “Devil’s staircases” in the system is shown in Fig. 4, where separate curve, corresponded to its value of varying parameters in the unit interval $0\leq a=c \leq 1$, is presented by its color. In the borders of the parameters range the staircase turns into the piece of straight line.

There are examples of top tree structures for another ordering method, called as TRD (“Ternary to Decimal – along Reverse time”), in Fig. 5. For this method the ordering function is defined by $G(n)=3^{n-1}$. This results in inverting the position order in symbolic dynamics (so as high-order and low-order digits trade places) in comparison with TDD method.

Fig. 3. Top tree structures, obtained by TDD method, applied to reverse iterations in (2); $N=12$. 
Fig. 4. Family of top tree structures (“Devil’s staircases”); N=8, 0≤a=c≤1.

a) a=1/3, b=1, c=2/3, d=0; dF ≈1.9.

b) a=c=0.5, b=1, d=0; dF ≈1.9.

Fig. 5. Top tree structures, obtained by TRD method, applied to reverse iterations in (2); N=10.

As in the case of two-valued nonlinearities of maps in reverse time, here also a number of other ordering methods can be proposed, which will have its own ordering function or, may be, form of conversion (4).

4. Conclusions

The fractal distribution of reverse iteration points of three-segment piecewise-linear map is identified. The ordering principle of these points is developed for three-valued systems. The structures, obtained by using of proposed two ordering methods (TDD and TRD), have self-similar fractal forms.

Comparing two- and tree-valued systems in reverse time gives us a deduction about its qualitatively analogous behaviors. However, formed fractal sets are different: the considered here structures have three self-similar parts, but in two-valued systems – two self-similar parts and there are some other fine distinctions. "Devil’s staircase", that is the only non-trivial fractal structure representing the solid manifold in the system, is not observed in two-valued systems.

The TDD method is proposed here as approach to investigation of fractal phenomena in many-valued maps in reverse time. The TRD method results in complicated random-like structures, which can be used for cryptosystem construction.

Fractal phenomena in reverse time have to be observed in discrete maps with any order of many-valuedness, but this generalization would be proved in further researches. Application problems are also not considered in this paper in detail. However, concerning to choice of many-valuedness order in cryptosystem, we can suppose that the higher the order of many-valuedness, the higher the security of encrypted information, but the higher computational effort is required.

References

Applications of High-Dimensional Cat Map and Its Mixing Nature

H.S. Kwok and Wallace Tang

Department of Electronic Engineering,
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
Email: hskwok, kstang@ee.cityu.edu.hk

Abstract—The use of high-dimensional Cat maps in cryptographic applications has been widely explored. This paper is the first attempt to investigate the mixing nature of these maps by the use of a statistical approach. The study also provides some guidelines in the determination of the parameters, such as dimensions and number of iterations, so that sufficient mixing effect can be obtained from these maps.

1. Introduction

Cat map is one of the chaotic maps commonly used in cryptographic applications [1, 2, 3, 4]. The original Cat map is a two-dimensional map which is described in [5]. It has been further extended to three dimensions [1], and later a general form of high-dimensions [3].

The interest of using high-dimensional Cat map is initiated by its simplicity and chaotic nature, in particular its mixing nature. However, a detail study of this mixing nature has never been given, probably due to the complexity caused by the high dimensionality.

In [6], the behavior of one dimensional discrete-time chaotic maps has been analyzed and described by statistical models. In this paper, the similar concept is applied to manifest the mixing property, together with the rate of mixing, of the high-dimensional Cat maps.

The organization of this paper is as follow. In Sect. 2, the high-dimensional Cat map is formulated and some typical applications are explained briefly. The statistical analysis of this map is then given in Sect. 3. Finally, conclusion is drawn in Sect. 4.

2. High-dimensional Cat Map

An $m$-dimensional Cat map [3, 4] is expressed as below:

$$
\begin{align*}
  x_i(k+1) &= \left[ \begin{array}{c} x_1(k) \\ x_2(k) \\ \vdots \\ x_m(k) \end{array} \right] = A \left[ \begin{array}{c} x_1(k) \\ x_2(k) \\ \vdots \\ x_m(k) \end{array} \right] \mod 1 \\
  \text{where } A &= A_{12}A_{13} \cdots A_{1m}A_{23} \cdots A_{2m} \cdots A_{(m-1)m} \text{ is a } m \times m \text{ constant matrix constructed by multiplying } m \times m \text{ matrices,}
\end{align*}
$$

$A_j$, each one mixing the $i$-th and $j$-th dimensions as in Cat map. The matrix, $A$, can be expressed as follows:

$$
A = \left[ \begin{array}{cccccc}
1 & a_{12} & 0 & \cdots & 0 & 1 \\
1 & 0 & a_{13} & \cdots & 0 & \cdots \\
1 & 0 & 0 & \cdots & a_{1m} & \cdots \\
0 & 0 & \cdots & 0 & 1 & \cdots \\
0 & 0 & \cdots & 0 & 0 & 1 \\
0 & 0 & \cdots & 0 & 0 & 0 \\
0 & 0 & \cdots & 0 & 0 & 0
\end{array} \right]
$$

where $0 \leq a_{ij}, b_{ij} \leq 1$.

Due to its chaotic natures, this high-dimensional Cat map has been used for different applications. Two of them are briefly explained in the followings.

2.1. Message Digestion

In [3], a cryptographic hash function, $h_c$, based on 10-dimensional Cat map has been designed. A hash function is a one-way function which should be efficiency and collision resistant. It can also generate a fixed length message digest value for any input with arbitrary length, as shown in Fig. 1.

The operational procedure is as follows. Message is divided into blocks of $n \times 32$ bits ($n=6$) and each block, used as the initial condition, is processed by iterating the 10-dimensional Cat map implemented in 32-bit
environment. The remaining states are either obtained from the initial buffers or by iterating the previous block. Finally, an 128-bit message digest is then obtained after processing all the blocks.

Due to the mixing feature of the map, an avalanche effect is observed. Even for a single bit difference in the messages, a large bit difference, with the mean close to the half of the length of the message digest, is obtained as shown in Fig. 2. A very fast speed can be achieved, for example, its throughput rate is about 1.5 times higher than MD5 [7].

The map in Eq. (2) can also be used for image encryption based on a stream cipher framework. Images are fed as a data stream and masked with keystream generated by a skewed Tent map and the high-dimensional Cat map, as shown in Fig. 3. In this design, the high-dimensional Cat map is served as a post-processing function to further mix up the bit stream generated from the Tent map. It is proved that it can enhance the randomness of the keystream and provide favorable results in all the statistical tests. An illustrative example, showing original and encrypted images, is given in Fig. 4. The encrypted image is appeared to be random and statistically uncorrelated to the original image. This algorithm also gives a desirable encryption speed when comparing with some existing algorithms.

### 3. Statistical Analysis of Chaotic Dynamics

From the applications described before, it can be observed that the mixing nature of Eq. (2) is the key for success. However, due to the high dimensionality, it is difficult to provide a rigorous proof for it. In addition, the optimal dimensions and iterations for providing sufficient mixing characteristics are also difficult to be determined.

In this paper, the concept described in [6] is used as the methodology to tackle with the above two questions. Our approach is to perform a statistical study on Eq. (2), by considering it as a black box with single input and single output.

Given that a set of normally distributed initial conditions is chosen as the initial values of \( x_1 \) of Eq. (2) while the initial conditions of \( x_2, x_3, \ldots, x_m \) and the parameters \( a_{ij} \) and \( b_{ij} \) are randomly chosen from \([0,1]\). After a fixed number of iterations (denoted as \( K \)), the final values of state \( x_1 \) are categorized into \( b \) non-overlapping bins, \( Y_i = \{ (i-1)/b, i/b \} \) where \( i = 1, 2, \ldots, b \). In our tests, \( b \) is always set to 100. The frequency count of each bin can be computed by:

\[
\frac{\# \{ x_1(n) \in Y_i, n = 1,2,\ldots,N \}}{N}
\]  

(3)

where \# denotes the cardinality of a set, and \( N \) is the size of the initial values set, which is assumed to be 20,000.

### 3.1. Testing Parameters

Two sets of initial conditions, \( S_1 \) and \( S_2 \), are randomly generated as depicted in Fig. 5, with mean 0.4 and 0.8 and standard deviation (SD) 0.1 and 0.001, respectively.
By studying the frequency count distribution histograms, it can manifest the mixing nature of the high-dimensional Cat map expressed in Eq. (2).

### 3.2. Mixing Nature of High-Dimensional Cat Map

Resultant histograms with initial condition sets $S_1$ and $S_2$ are obtained as shown in Fig. 6 and 7, respectively, using a 5-dimensional Cat map with different numbers of iterations ($K = 1, 10, 20$).

From Figs. 6 and 7, there are some interesting features observed:
1. The original set of initial values is spread all over the entire range.
2. The distributions of the histograms are almost invariant after a few iterations, and are independent of the initial values.

The independency of initial value is further confirmed by varying the value of the mean and the standard derivation (SD) of $S_2$, and the results are shown in Fig. 8.

To have a good mixing effect, the final histogram should be uniform without bias. This can be justified by a Chi-squared test [8]. The hypothesis of the test is that the obtained frequency histogram after $K$ iterations is normally distributed over the phase space.

The chi-squared value $\chi^2$ in each case is compared with a reference value, $\chi^2_{a,df} = 123.23$, according to the level of significance $\alpha=0.05$ and the degree of freedom $df=b-1=99$. If $\chi^2 < \chi^2_{a,df}$, the histogram is uniform, implying that a desirable mixing effect is obtained. The results with different dimensions and number of iterations are tabulated in Table 1 (underlined data indicates $\chi^2 < \chi^2_{a,df}$).
It can be noticed that
1. The mixing quality is dependent on the dimension of the Cat Map, the higher the better.
2. The mixing quality is also related with the number of iterations. With a rough definition of the term “mixing speed”, we can say that the higher the dimension, the higher is the mixing speed.

### 3.3. The Effect of Dimensions and Iterations

From the previous section, it is noticed that the mixing speed and the quality are both dependent on the dimensions of the map. To further study their effects, a sequence of simulations is carried out with different Cat maps with different number of iterations.

Instead of showing the histogram of each test, its Chi-squared value $\chi^2$ is plotted in Fig. 9. Each vertical bar represents the mean Chi-squared value from a hundred sets of $S_i$, each with a random chosen mean and having SD equal to 0.001. With such a small SD, the initial points will be limited in one or two bins. If the mixing effect is sufficient, a uniform frequency histogram will be obtained, which can also be reflected by the value of $\chi^2$.

For the ease of referencing, a horizontal plane indicating $\chi^2_{\alpha,df}$ is given. Any bar above this reference plane implies that the histogram is not uniform, or the mixing effect is not good enough.

<table>
<thead>
<tr>
<th>Dim.</th>
<th>$K=1$</th>
<th>$K=10$</th>
<th>$K=20$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>169840.72</td>
<td>50670.65</td>
<td>20738.67</td>
</tr>
<tr>
<td>3</td>
<td>179346.79</td>
<td>1944.62</td>
<td>213.90</td>
</tr>
<tr>
<td>4</td>
<td>183508.05</td>
<td>136.79</td>
<td>106.20</td>
</tr>
<tr>
<td>5</td>
<td>172629.90</td>
<td>107.16</td>
<td>109.77</td>
</tr>
</tbody>
</table>

Table 1. Mean of $\chi^2$ from 100 sets of initial values with random mean, and SD=0.001 (same as $S_2$).

The mentioned features of the high-dimensional Cat map are further confirmed with this plot. By considering a fixed number of iterations, such as $K=2$, it can be noticed that $\chi^2$ decreases with the dimensions, implying that a better mixing nature is obtained for a higher dimension. From Fig. 9, it is also possible to select the appropriate dimension and the corresponding iteration so that a sufficient mixing effect can be achieved. For example, a 4-dimensional Cat map will need at least 11 iterations to achieve the nice mixing effect, while 6 iterations are needed if the dimension is five.

### 4. Conclusions

In this paper, the applications of high-dimensional Cat map are briefly introduced. With the use of statistical analysis, it is discovered that the high-dimensional Cat map is a transformation, possessing mixing property. Moreover, some guidelines for the choice of the dimension and the number of iterations are found, which are important for the future applications.

### Acknowledgments

The work described in this paper was fully supported by a grant from the Research Grants Council of the HKSAR, China [CityU 120205].

### References


On Some Properties of Chaotic Maps Implemented in Finite-State Machines

Mieczysław Jessa

Institute of Electronics and Telecommunications, Poznan University of Technology
ul. Piotrowo 3a, 60-965 Poznan, Poland
E-mail: mjessa@et.put.poznan.pl

Abstract—Relations between properties of orbits of finite-state implementations of one-dimensional chaotic maps and results known from the number theory were discussed. A new theorem and a corollary were proposed, relating the problem of shadowing of chaotic orbits with the ability of a chaotic map for generating each sequence of symbols from a set of uncountably many sequences of symbols.

1. Introduction

Finite-state implementation of chaotic maps has played the key role in the understanding of the chaotic dynamics. It is also important in practical applications because of many advantages of digital circuits, compared with analogue ones. Using chaotic maps and finite-state machines, we can generate the same signals in geographically separated places for different software or hardware platforms, and without sending a synchronization signal. Generators of different signals, prepared as software, can be easily integrated with existing digital systems, which lowers the cost of developing new solutions.

The basic problem arising in finite-state implementation of chaotic maps are relations between the properties of periodic cycles obtained after implementing a chaotic map in a finite-state machine and the properties of periodic and non-periodic orbits of the same map. Another problem concerns relations between the results of simulating chaotic maps and facts known from the number theory. The first of the problems is discussed in several papers (e.g. in [1], [2], [5], [7], [8]) but relations between the properties of periodic cycles of the implemented chaotic maps and facts known from the number theory are not studied sufficiently. The aim of the paper is to study these relations and to propose such approaches to the problem of implementing one-dimensional (1D) chaotic maps in digital circuits that may offer new applications of the maps, compared to solutions known from the number theory. It is also shown that some properties of 1D chaotic maps are preserved in digital circuits when the original map is capable for generating each sequence of symbols from the set of uncountably many sequences of symbols.

2. Computer Simulations of Chaotic Maps

Computer simulations of chaotic maps are the basic tool of chaotic dynamics. However, they have to be evaluated carefully when studying the long-term behavior of trajectory points of the maps. Assuming that continuous variable \( x \in [0,1) \) of certain chaotic 1D map \( f \) is turned into \( N \)-state variable \( \tilde{x} \), we obtain [2]

\[ \tilde{x} = \beta \cdot b^{-l}, \quad \beta = 1,2,\ldots,b^l, \]

where \( b \) is the base of a number system used in computations and \( l \) is the number of digits encoding numbers in this system. In digital systems \( b \) is equal to 2 and \( l \) is the number of bits encoding single \( \tilde{x} \). To each state \( \tilde{x}_{n+1}, n=1,2,\ldots \) as another iterate, state \( \tilde{x}_n \) is assigned, where \( \tilde{x}_n \leq x_n \) and \( x_n \) is the iterate of chaotic map \( f \) [2]. Condition \( \tilde{x}_n \leq x_n \) concerns both the truncation and rounding operations used in finite-state machines. It is known that the iteration process may amplify these errors to the point where the obtained orbits are not true orbits for given initial conditions. Due to the errors, most of the orbits terminate in periodical orbits with length

\[ L \sim e^{-d/2}, \]

where \( e \) is the size of a cell of the discretized phase space and \( d \) is the correlation dimension of the chaotic attractor. The conclusion concerns all maps with fractal chaotic attractor [5]. In the case of some 1D maps, the maximal length \( L_{\text{max}} \) of orbits is

\[ L_{\text{max}} \sim \sqrt{N}, \]

where \( N = 2^l \) is the number of phase cells of the discretized phase space [1]. Because the precise values for \( L \) and \( L_{\text{max}} \) are unknown, the use of such maps in many practical applications is limited. On the other hand, it can be exploited in studying chaotic dynamics because the true orbits stay near the false orbits obtained in a finite-state machine [4], [7], [8]. In such case we say that a true orbit shadows the false orbit. Anosov and Bowen proved the shadowing results for hyperbolic maps on a differential manifold. Sauer and York have generalized the approach of Anosov and Bowen on systems which are not necessarily hyperbolic. The case of 1D maps was considered by Chow and Palmer. The application of the shadowing to many well-known maps has to be considered very carefully. It concerns for example the sawtooth map

\[ S_\lambda(x) = \lambda \cdot x \mod 1 \]

and the tent-like map
The described case occurs for all continuous variables \( x \) or is mapped into rational numbers. It seems to be difficult to formulate a general answer, but studying the properties of the same chaotic maps, one can notice that rational numbers are always equivalent in the set of integer numbers (Fig. 1). If \( q_n \) is constant during all iterations, we obtain

\[
\tilde{X}_n = \tilde{X}_{n-1} \mod 1,
\]

where \( \lambda = k + i \cdot b^{-1} \), \( k=1,2,3,\ldots, i=0,\pm 1,\pm 2,\ldots, \pm b^{-1} \). The obtained trajectory \( \{\tilde{X}_n\} \) is the same as sequence

\[
\tilde{X}_n = (p_n)/(b^j),
\]

and \( 0 \leq p_n < b^j \), \( n=1,2,3,\ldots \). If \( i=0 \) or \( i=\pm 1 \), then term \( i \cdot b^{-1} \cdot p_{n-1} \) is a fraction with values from open interval \((-1,1)\) and has no influence on the value of \( p_n \). Therefore, formula (9) can be replaced by equation

\[
p_n = (k \cdot p_{n-1} + i \cdot b^{-1} \cdot p_{n-1}) \mod b^j.
\]

It was obtained that the properties of orbits \( \{\tilde{X}_n\} \) of a chaotic map \( f \) considered for \( \lambda = k + i \cdot b^{-1} \), where \( i=0 \) or \( i=\pm 1 \) are the same as properties of sequences produced by the multiplicative pseudorandom number generator, well-known in the number theory. Assuming the truth of the shadowing, we have also that for the specified values of parameters the properties of non-periodic orbits of the sawtooth map, found in simulation experiments, are determined by the properties of sequences of integer numbers produced by known pseudorandom number generator. The same conclusion concerns the tent-like map with \( \lambda = k + i \cdot b^{-1} \), where \( i=0 \) or \( i=\pm 1 \). It results from the equation relating the sawtooth and the tent-like map. From (4) and (5) it is that

\[
x_n = \begin{cases} 
S_\lambda(x_{n-1}) & \text{for even} \\
1 - S_\lambda(x_{n-1}) & \text{for odd}
\end{cases}
\]

Because the properties of the sawtooth map, found in a simulation experiment, for integer values of \( \lambda \) are sufficient to predict the properties of the same map for rational values of \( \lambda \), where \( \lambda > 1 \), we may assume that they are also fully determined by the properties of the multiplicative pseudorandom number generator, when the truth of the shadowing theorem was assumed. Let us emphasize that the conclusions do not concern all properties of the sawtooth or the tent-like map but only such of them that can be derived from simulation experiments performed in finite-state machines.

The example presented discourages searching practical applications of the mentioned maps in finite-state machines because the properties of the multiplicative pseudorandom generator are well-known. On the other hand, it can be noticed that results of simulating 1D chaotic maps with the use of the floating point arithmetic are very specific because they are obtained only for a subset of rational numbers. Therefore, a more general problem should be considered. We may ask if applying a 1D chaotic map in a finite-state machine new results can be obtained or if they can be only a repetition or a modification of facts or properties known in the theory of integer numbers. It seems to be difficult to formulate a general answer, but studying the properties of the same

\[
T_\lambda(x) = \left( (-1)^{[\frac{x}{\lambda}]_{\mod 1}} \cdot \lambda \cdot x \right) \mod 1
\]

with real values of parameter \( \lambda \). These maps, implemented in the floating point arithmetic, do not generate true orbits for a large number of iterations. When the parameter is integer and even, the shadowing fails because for large \( n \) the orbits terminate in sequences of zeros for all initial points. If the parameter is odd or rational, the orbits are still false, but the shadowing gives true values of some parameters characterizing the chaotic attractor (e.g. the KS entropy, the distribution of trajectory points).
map in the set of all rational numbers \( \hat{x}_n = (p_n)/(q_n) \) available in a finite-state machine, we may try to find examples of new results, useful in practical applications. To study the properties of \( f \) in the set of all rational numbers \( \hat{x}_n = (p_n)/(q_n) \), numbers \( p_n \) and \( q_n \) have to be encoded separately. Let us consider two cases. In the first case we assume that \( q_n \) is constant during all iterations and equal to e.g. \( q_0 \) which is not the second or a higher power of an integer number. In the second case \( q_n \) may vary during iterations of an initial state. For the latter one the formula assigning 1:1 integer numbers \( n_n \) to rational numbers \( \hat{x}_n = (p_n)/(q_n) \) may be different for different iterations. Consequently, we cannot find a formula for \( g \), generating sequences of integer numbers with the same properties as the sequences of rational numbers obtained as the result of simulating chaotic map \( f \) in a finite-state machine. In this case, the results of theory of integer numbers cannot be simply applied to studying the properties of \( f \). It may open new possibilities for finding applications of chaotic \( f \), which cannot be derived directly from the number theory.

The choice of constant \( q_n \), which is not the second or a higher power of an integer number, simplifies computations but enables the introduction of map \( g \). If \( q_0 = b^j \), then numbers \( \hat{x}_n = (p_n)/(q_0) \) have periodic expansions in a number system with base \( b \). The periodic segment starts from the first or a later digit of number \( \hat{x}_n \) and the formula for map \( g \) takes the form:

\[
p_n = g(p_{n-1}) = q_0 \cdot f(q_0^{-1} \cdot p_{n-1})
\]

The existence of the formula for \( g \) generating sequences of integer numbers with known properties, determining also the properties of orbits \( \{ \hat{x}_n \} \) of map \( f \), does not exclude the introduction of new applications exploiting \( f \) implemented in a finite-state machine. Let us illustrate it with an example.

**Example 2**

It is known that map \( S_\lambda \) can be used for generating binary sequences. The sequences can be generated from formula

\[
\left\{
\begin{align*}
    a_n &= \lfloor \alpha \cdot x_n \rfloor \\
    x_n &= S_\lambda(x_{n-1})
\end{align*}
\right.
\quad n = 1, 2, 3, ..., (13)
\]

where \( \lambda > 1 \) is integer and \( \alpha = 2 \) is the number of disjoint subintervals \( I_j \), \( j = 0, 1 \) obtained as the result of partitioning the unit interval \( I = \{0, 1\} \). To avoid situations when sequences of symbols \( \{a_n\} \) are the same as digits encoding \( x_0 \) written in the fixed point arithmetic with base \( \lambda \) or when some groups of symbols of sequence \( \{a_n\} \) are the same as digits encoding \( x_0 \), we assume that subintervals \( I_j \) do not cover with subintervals at which map \( S_\lambda \) is monotonic, and that numbers \( \alpha \) and \( \lambda \) are not the powers of the same integer number [6]. Implementing map \( S_\lambda \) in a finite-state machine, we obtain

\[
\left\{
\begin{align*}
    a_n &= \left\lfloor \frac{2}{q_0} \cdot p_n \right\rfloor \\
    p_n &= (\lambda \cdot p_{n-1}) \mod q_0, \quad n = 1, 2, ...
\end{align*}
\right.
\]

(14)

where \( q_0 = 2^j \) for \( \hat{x}_0 \) encoded by a fixed number of digits and \( q_0 \neq 2^j \) for \( \hat{x}_0 \) determined by integer numbers \( p_n \) and \( q_0 \) encoded separately. In both cases, symbol \( a_n \) is encoded by higher-order digits of number \( \hat{x}_n \) obtained through dividing numbers \( p_n \) by \( q_0 \). The difference is that for \( q_0 = 2^j \) higher-order digits of \( \hat{x}_n \) are the same as higher-order digits of \( p_n \) and for \( q_0 \neq 2^j \) such situation does not occur. Consequently, when \( q_0 = 2^j \) successive symbols can be computed from already generated symbols because the Boyar’s algorithm of inferring sequences \( \{a_n\} \) can be used in such case [3]. If \( q_0 \neq 2^j \), the algorithm fails because we do not know any digit of integer numbers \( p_n \) produced by generator (14) [6].

The example presented shows that studying the properties of chaotic maps in the whole set of rational numbers available in a finite-state machine, we may obtain new results even when \( q_n \) is constant and equal to e.g. \( q_0 \). Consequently, new applications of \( f \) can be searched both when \( h \) is unknown or when \( q_0 \) is fixed. In the first case, we can prepare a software for computing successive \( q_n \) or we can apply different approaches ensuring unknown \( h \) and \( g \), e.g. the interval arithmetic. When \( q_0 \) is fixed, the computations are simpler but we have to verify carefully if the results of simulations or applications of chaotic maps in finite-state machines are not a repetition or a simple modification of facts or properties known from the number theory.

4. Properties of 1D Chaotic Maps Preserved in Finite-State Machines

Results of computer simulations of chaotic maps are commonly used to study the properties of these maps. It is worth noticing that such approach is not obvious because periodic orbits considered inside their periods are not the same as parts of non-periodic orbits. Good examples are the sawtooth map and the tent-like map. Assuming rational values of \( \lambda \), we have that all non-periodic orbits are composed of points which are irrational numbers. No rational number can be an element of a non-periodic trajectory. Implementing a chaotic map in a finite-state machine, we observe only periodic cycles which contains only rational numbers. No part of a periodic cycle can be a part of a non-periodic one. It may suggest that investigating the periodic cycles, we cannot obtain proper knowledge about the properties of non-periodic orbits and about the whole attractor containing uncountably many
non-periodic orbits and countably many periodic ones. On
the other hand, simulation experiments and theoretical
results prove that computer modeling of chaotic orbits
may give true information about the properties of non-
periodic orbits. These observations can be supported by
the following new theorem:

**Theorem**

Let us assume that a one-dimensional chaotic map $f$ can
generate each sequence of symbols. Symbols are assigned
to subintervals of width $\delta$, obtained as the result of
partitioning chaotic attractor $X$ of map $f$ into a finite
number of disjoint subintervals $I_j$, i.e.

$$
X = \bigcup_j I_j \quad I_j \cap I_k = \emptyset, \quad i \neq j
$$

$$
a_n = j \quad \text{for} \quad x_n \in I_j
$$

$$
x_n = f(x_{n-1}) \quad n = 1, 2, 3, \ldots.
$$

(15)

Then, there exists a non-periodic orbit which shadows the
false orbit in time interval shorter than the period of
sequence $\{a_n\}$ obtained in a finite-state machine with
$N \geq \delta^{-1}$ states. The false orbit stays within the shadow
of width $\delta$ of the true orbit.

**Proof**

If a chaotic map $f$ can generate each sequence of symbols
it can generate infinitely many non-periodic sequences and
infinitely many periodic ones. The periodic sequences
$\{a_n\}$ with finite periods can be obtained in a finite-state
machine with $N \geq \delta^{-1}$ states. The periods of $\{a_n\}$ depend
on $f$ and on $N$. In such case any part of a periodic sequence
$\{a_n\}$, shorter than the period, is a part of a non-periodic
sequence $\{a_n'\}$, obtained with the use of the same map but
in the set of real numbers. Thus, if a false orbit obtained in
a finite-state machine generates sequence $\{a_n\}$, then there
exists a non-periodic orbit which produces a non-periodic
sequence $\{a_n'\}$ which is the same as sequence $\{a_n\}$ in
observation interval $I_o$ shorter than the period $T_o$ of
sequence $\{a_n\}$. It implies that a non-periodic orbit
$\{a_n'\}$, shadows the false orbit generating $\{a_n\}$ for
$t_o < t_p$, and the false orbit stays within the shadow of
width $\delta$ of the true orbit.

Exploiting the fact that most of the orbits terminate in
periodical orbits with length (2), we may formulate the
following corollary:

**Corollary**

If a one-dimensional chaotic map $f$ can generate each
sequence of symbols and $\delta$ is the width of intervals $I_j$ with
assigned symbols $a_m$ defined by (15), then there exists a
true orbit which shadows for $L \sim e^{-d/2}$ iterations the
false orbit generating periodic $\{a_n\}$ in a finite-state
machine with $N$ phase cells of size $\epsilon = \delta$, when it is
assumed that the period of $\{a_n\}$ is equal to the period of
the false orbit.

The corollary concerns one-dimensional maps but it
relates the shadowing with the correlation dimension of
the chaotic attractor. For typical maps it is that $d=1$ [5]
and the corollary is similar to that obtained in papers [4],
[7] and [8] but it includes the case of $S(x)$ and $T_p(x)$ with
even $\lambda$, described in Section 2. The novelty of the
proposed theorem and corollary is that it relates the
shadowing with the ability of a chaotic map (e.g. the
sawtooth map) for generating each sequence of symbols
from uncountably many existing sequences of symbols.
The sequences are obtained as the result of partitioning
the chaotic attractor into disjoint subsets covering the
whole attractor.

5. Conclusions

The properties of chaotic maps implemented in a finite-
state machine were discussed. It was shown that finite-
state implementation of chaotic 1D maps may lead to
results which are known in the number theory or they are
new from the perspective of the same theory, depending
on the applied method of coding of trajectory points in
finite-state machines. A new approach to the problem of
shadowing of the orbits of chaotic maps was proposed. It
relates the shadowing with the ability of a chaotic map to
generate each sequence of symbols. The basic task for the
nearest future is the generalization of the presented results
on multi-dimensional chaotic maps.

References

discretization on the long-time behavior of dynamical

with finite-state machines,” Phys. Rev. A, A34, pp. 4460-
4463, 1986.

1, pp. 177-184, 1989.

of orbits of dynamical systems: the one-dimensional case,” J.
Dynamics and Differential Equations, vol. 3, pp. 361-379,

periodicity and the correlation dimension of chaotic

generated by means of the sawtooth chaotic map,” IEEE
Transactions on Circuits and Systems – I: Regular Papers,

pseudorandomness,” in Chaos Control, Chen G., Yu X., eds,

trajectories for the computer simulation of dynamical
Higher Order Dependency of Chaotic Maps

A.J. Larwance† and T. Papamarkou†
†Department of Statistics, University of Warwick
Coventry, CV4 7AL, UK
Email: A.J.Lawrance@warwick.ac.uk, T.Papamarkou@warwick.ac.uk

Abstract—Some higher-order statistical dependency aspects of chaotic maps are presented. The autocorrelation function (ACF) of the mean-adjusted squares, termed the quadratic autocorrelation function, is used to access non-linear dependence of the maps under consideration. A simple analytical expression for the quadratic ACF has been found in the case of fully stretching piece-wise linear maps. A minimum bit energy criterion from chaos communications is used to motivate choosing maps with strong negative quadratic autocorrelation. A particular map in this class, a so-called deformed circular map, is derived which performs better than other well-known chaotic maps when used for spreading sequences in chaotic shift-key communication systems.

Key Words—quadratic autocorrelation, piece-wise linear maps, circular maps, negative statistical dependency, chaos shift-keying communication systems

1. Introduction

The linear statistical dependence of fully stretching piece-wise linear maps has been explored in Baranovski & Daems [2], where an explicit expression for their ACF was found, and in Baranovski & Lawrance [1]. This work is extended here by further analyzing the non-linear dynamics of these maps. More specifically, their mean-adjusted quadratic ACF has been found. The result is interesting in the sense that certain collections of fully stretching piece-wise linear maps share the same quadratic ACF.

The intention of minimizing the bit energy of a chaotic communication system, e.g., chaos-shift keying, Lawrance and Ohama [5], led to the formulation of a criterion for assessing the performance of the chaotic map of the system (see Yao [7]). The criterion focuses on the non-linear dependency of the spreading sequence and specifies the form of the quadratic ACF of an effective map.

Finally, the first-order circular map is deformed to decrease its negative quadratic dependency to nearer the absolute lower bound, and so make it nearer to optimal for spreading in chaos-based communication systems.

2. Dependence for Piecewise Linear Maps

2.1. Four Well-Known Maps

We start with the two-branch equi-distributed piece-wise linear maps, namely with the tent and valley map, the positive and negative Bernoulli shift map. They fulfill the equi-distributed property (Khoda & Tsuneda [3], Khoda & Tsuneda [4]) and they are symmetric about the line $y = \frac{1}{2}$, as it can be seen in Figure 1.

![Figure 1: The four two-branch equi-distributed maps defined on [0,1].](image)

A perhaps surprising dependence property of the four maps is that they share the same adjusted quadratic ACF, whereas their linear ACFs differ (Table 1).

<table>
<thead>
<tr>
<th>Map</th>
<th>Linear ACF</th>
<th>Quadratic ACF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tent</td>
<td>0</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>Valley</td>
<td>0</td>
<td>$\frac{1}{7}$</td>
</tr>
<tr>
<td>Positive Bernoulli</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>Negative Bernoulli</td>
<td>$\frac{1}{2}$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

Table 1: The linear and quadratic ACF of the four maps.
2.2. Two-Branch Piecewise Linear Maps

We now consider the more general collection of two-branch piece-wise linear maps. This particular family embraces maps whose monotonicity switches direction at some point \( r \in (c, d) \) of their domain \([c, d]\). We introduce the term non-centrality parameter to refer to point \( r \). The four equi-distributed maps previously encountered can be seen as special cases arising from the choice \( r = 0.5 \) of the non-centrality parameter.

Figure 2 exemplifies the collection of non-central two-branch maps and Table 2 provides their statistical dependence.

![Piece-wise linear examples of non-central two-branch maps.](image)

**Table 2:** The linear and quadratic ACF of the four non-central maps

<table>
<thead>
<tr>
<th>Map</th>
<th>Linear ACF</th>
<th>Quadratic ACF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tent</td>
<td>((2r - 1)^2)</td>
<td>((3r^2 - 3r + 1)^2)</td>
</tr>
<tr>
<td>Valley</td>
<td>((1 - 2r)^2)</td>
<td>((3r^2 - 3r + 1)^2)</td>
</tr>
<tr>
<td>Positive Bernoulli</td>
<td>((2r^2 - 2r + 1)^2)</td>
<td>((3r^2 - 3r + 1)^2)</td>
</tr>
<tr>
<td>Negative Bernoulli</td>
<td>((-2r^2 + 2r - 1)^2)</td>
<td>((3r^2 - 3r + 1)^2)</td>
</tr>
</tbody>
</table>

It is worth mentioning that the four non-central maps of Table 2 are uniformly \( U(0, 1) \) distributed regardless of the non-centrality parameter \( r \). But the most intriguing fact is that they share the same adjusted quadratic autocorrelation function, provided that they all have the same non-centrality parameter \( r \); this point is still open to intuitive explanation.

2.3. Fully Stretching Piece-wise Linear Maps

The highest level of our generalization involves fully stretching piece-wise linear maps \( \tau : [0, 1] \to [0, 1] \) with more than two branches. We assume that \( \tau \) is composed of \( k + 1 \) linear branches \( \tau_i(x) = \lambda_i x + \mu_i \), i.e. that \( \tau \) has \( k \) non-centrality parameters \( r_i, \ i \in \{1, 2, \ldots, k\} \), such that \( 0 < r_0 < r_1 < r_2 < \ldots < r_k < r_{k+1} = 1 \).

It can be shown that the uniform \( U(0, 1) \) satisfies the Perron-Frobenius equation, that is \( U(0, 1) \) can be adopted as the invariant density of the chaotic sequence produced by the map \( \tau \).

The linear ACF of fully stretching piece-wise linear maps is known (see Baranovski & Daems [2]), and is neatly given by

\[
Corr(X_t, X_{t+s}) = \left( \frac{1}{\lambda_i} \right)^s, \ s \in \mathbb{N}, \quad (1)
\]

where \( \lambda_i \) is the slope of the \( i \)-th branch \( \tau_i \) of \( \tau \).

We report that the adjusted quadratic ACF of fully stretching piece-wise linear maps is correspondingly given by

\[
Corr\left(\left(X_t - \frac{1}{2}\right)^2, \left(X_{t+s} - \frac{1}{2}\right)^2\right) = \left( \frac{1}{\lambda_i} \right)^s, \ s \in \mathbb{N}. \quad (2)
\]

An alternative expression for (2), in terms of the non-centrality parameters \( r_i, \ i \in \{1, 2, \ldots, k\} \), of \( \tau \) is provided as:

\[
Corr\left(\left(X_t - \frac{1}{2}\right)^2, \left(X_{t+s} - \frac{1}{2}\right)^2\right) = \left( \sum_{i=1}^{k+1} \frac{1}{\lambda_i^2} \right)^s. \quad (3)
\]

We point out the conclusions which can be drawn from our work so far:

- Equation (3) generalizes and clarifies the property we came across in Tables 1 and 2; all fully stretching piece-wise linear maps with the same non-centrality parameters \( r_i, \ i \in \{1, 2, \ldots, k\} \), share the same adjusted quadratic ACF. On the other hand, they do not exhibit identical linear dependence, since their linear autocorrelations do not coincide.

- A comparison between equations (1) and (2) justifies the choice of adjusting the quadratic ACF for the mean; the resemblance between the two equations suggests that the adjusted quadratic ACF is a natural measure of non-linear dependence, principally because it is not affected by the mean.

- Finally, it can be deduced from equation (3) that the adjusted quadratic autocorrelations of any lag of any fully stretching piece-wise linear map is positive. However, this is precisely not wanted for chaos-based spreading in communication systems, as we will be seen in section 3.

3. Criterion for Optimality

The bit energy of several binary communication systems, such as chaos shift-keying, involving chaotic spreading sequences, is a function of the mean-adjusted sum of squares
of the spreading, say
\[ \sum_{i=1}^{N} (X_i - \mu)^2. \]

See Lawrance & Ohama [5], for instance.

In such systems, the bit error rate can be shown to be minimised when the bit energy is constant at its minimum value. In particular, this implies that the variance of the bit energy should be zero and this is the starting point of the derivation of an optimum map \( \tau \), as set out in Yao [7].

First there is the routine general result
\[ \text{var} \left( \sum_{i=1}^{N} (X_i - \mu)^2 \right) = N \sigma^2_{|X_{\tau} - \mu|^2}, \]
where \( \sigma^2_{|X_{\tau} - \mu|^2} \) is the variance of the squares and \( \rho_{|X_{\tau} - \mu|^2}(i) \) is the lag \( i \) autocorrelation of the squares. This variance is clearly zero when
\[ 2 \sum_{i=1}^{N-1} \left( 1 - \frac{i}{N} \right) \rho_{|X_{\tau} - \mu|^2}(i) = -1. \]

Continuing in general from this result to specify the map \( \tau \) looks hopeless, and so in the next section it is usefully specialized.

4. Maps Consistent with the Optimality Criterion

4.1. An Optimal Circular Map

The smallest extent of spreading is \( N = 2 \) which when used in (5) gives the perfect correlation \( \rho_{|X_{\tau} - \mu|^2}(1) = -1 \) and more clearly \( \rho_{|X_{\tau} - \mu|^2,|\tau(X_{\tau} - \mu)|^2} = -1 \), showing that \( (X - \mu)^2 \) and \( (\tau(X - \mu))^2 \) are negatively linearly related. Assuming \( \mu = 0 \) without loss, this perfect negative correlation thus requires \( \tau(x)^2 = ax^2 + b \), \( a < 0 \), \(-1 \leq x \leq 1\), and hence if the square-root linear map is to be fully stretching over \((0, 1)\)
\[ \tau(\sqrt{x})^2 = x + 1, \quad -1 \leq x \leq 1. \]

This straight line cannot be a proper chaotic map and so dynamical straight line approximations are entertained. The simplest seems to be the negative sloped Bernoulli map illustrated in Figure 1, which suggests taking
\[ \tau(\sqrt{x})^2 = \begin{cases} -2x + 1, & 0 \leq x < \frac{1}{2} \\ -2x + 2, & \frac{1}{2} \leq x \leq 1 \end{cases}. \]

Proceeding with this map, possible forms of the map \( \tau(x) \) satisfying (7) are included in the formula
\[ \tau(\sqrt{x}) = \begin{cases} \pm \sqrt{2x + 1}, & 0 \leq x < \frac{1}{2} \\ \pm \sqrt{2x + 2}, & \frac{1}{2} \leq x \leq 1 \end{cases}. \]

Choosing the most natural possibility for \( \tau(x) \) gives the desired circular map
\[ \tau(x) = \begin{cases} -\sqrt{2x^2 + 2}, & -1 \leq x < -1/\sqrt{2} \\ \sqrt{2x^2 + 2}, & -1/\sqrt{2} \leq x < 1/\sqrt{2} \\ -\sqrt{2x^2 + 2}, & 1/\sqrt{2} \leq x \leq 1 \end{cases}, \]
which is illustrated in Figure 3. Yao [7] proves that the invariant distribution of this map has probability density function \( f(x) = |x|, \quad -1 \leq x \leq 1 \) and that it is linearly uncorrelated. We report that its quadratic autocorrelation function is
\[ \text{Corr}(X_{\tau}^2, X_{\tau+s}^2) = \left( -\frac{1}{2} \right)^s, \quad s \in \mathbb{N}, \]
still room for improvement of the lag 1 autocorrelation towards -1.

4.2. A Deformed Circular Map

In trying to reduce further the lag 1 quadratic ACF, the circular map is generalized by introducing a deforming parameter \( r \), which determines the range of each branch of the map. To be more specific, the deformed circular map is defined as:
\[ \tau(x) = \begin{cases} -\sqrt{(1-r)^{-1}x^2 + (1-r)^{-1}}, & -1 \leq x < -\sqrt{r} \\ \sqrt{r^{-1}x^2 + 1}, & -\sqrt{r} \leq x < \sqrt{r} \\ -\sqrt{(1-r)^{-1}x^2 + (1-r)^{-1}}, & \sqrt{r} \leq x \leq 1 \end{cases}. \]

This map has an invariant density \( f(x) \) given by
\[ f(x) = \begin{cases} -2(1-r)x, & -1 \leq x \leq 0 \\ 2rx, & 0 < x \leq 1 \end{cases}, \]
which satisfies the Perron-Frobenius equation, and is seen to generalize the invariant density \(|x|\) of the circular map. Moreover, Figure 4 helps us get a better idea about the shape of deformed circular maps by providing the plots of maps with deforming parameters \( r = 0.1 \) and \( r = 0.42 \) as well as their associated invariant densities.

![Figure 3: The circular map.](image-url)

![Figure 4: Two examples of deformed circular maps and their invariant densities.](image-url)
Although the derivation of an analytic expression for the adjusted quadratic ACF appears to be a cumbersome task, at least an explicit formula has been obtained for the first lag. The first lag can be expressed as a function of \( r \) only. Figure 5 provides the plot of lag 1 versus \( r \). As can be seen, the function is minimized for \( r = 0.42 \) and its minimum equals -0.722. In other words, we have found that the deformed circular map with deforming parameter \( r = 0.42 \) (Figure 4.(b)) is even better than the circular map (\( r = 0.5 \), Figure 3) from a communications perspective; it is closer to the requirements imposed by our criterion.

**Figure 5:** Plot of the first lag of the adjusted quadratic ACF of the deformed circular map versus its deforming parameter \( r \). The first lag takes its minimum value \(-0.722\) for \( r = 0.42 \).

Since this paper deals with work-in-progress, it is appropriate to point out that there is a further question as to how near -0.722 is to the theoretical lower bound. There is such a correlation lower bound over all bivariate distributions with specified marginal distribution. This is part of an area of distribution theory developed by Frechet (1951), see Ripley [6], for instance. What is needed is the distribution function of \( X^2 \) from (12) in this case, and the corresponding inverse function \( F^{-1}_{X^2}() \). The 'degenerate' bivariate distribution of minimum correlation is then that of the variables \( X^2 \) by the well known ‘inverse transformation’ of a uniform random variable \( U \), and because \( 1 - U \) is also a uniform random variable, the second also has the distribution of \( X^2 \). In the present context, this degenerate case of minimum correlation can be cast as

\[
X_{t+1} = F^{-1}_{X^2}(1 - F_{X^2}(X_t))
\]

(13)

and this would suggest that the map giving minimum correlation should approximate the curve

\[
\tau(x) = F^{-1}_{X^2}(1 - F_{X^2}(x))
\]

(14)

in a dynamical way.

**Acknowledgments**

We would like to thank the NOLTA2006 organizing committee for giving us the opportunity to present this work.

**References**


A short duration noise suppression method for speech signals by using a sparse signal representation

Makoto NAKASHIZUKA

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531 JAPAN
Email: nkszk@sys.es.osaka-u.ac.jp

Abstract—In this paper we propose a short duration noise suppression method based on a sparse signal representation. For speech and noise separation, we assumed that the speech is stationary within 20-40ms and the duration of the noise is shorter than this period. In our approach, a sparse representation is employed to separate the noise and speech by the difference of their time duration properties. For the sparse representation, a pair of DFT bases that support different time interval were employed to the sparse signal representation. The shorter and the longer DFT bases represent the noise and the speech respectively with a penalty of sparseness. In echoic environments the reverberation of the noises appears in the separated speech signals. In order to suppress the reverberation of the noise, we apply a spectrum subtraction to the separated speech. For the spectrum subtraction, we propose a power estimation method for the noise reverberation. In experiment, we apply the proposed method to noisy speech signals that are corrupted by noise bursts recorded in an echoic environment. We demonstrate that the proposed method can improve about 7-10dB in SNR of the noisy segments.

1. Introduction

Noise suppression for speech signals is one of major subjects in signal processing. Many noise suppression methods have been proposed[1][2]. Usually, the noise desired to be suppressed is assumed as a stationary signal. For example, the spectrum subtraction[1][2] that suppress the noise power spectrum in frequency domain requires estimation of the noise power. The noise is assumed to be stationary and its power is estimated in the interval where speeches are not active. However, for the short duration noises, such as impulsive, shot and burst noises, it is difficult to estimate the noise power spectrum. The noise suppression method that performs under the assumption of the noise stationarity fails to suppress the short duration noises. For suppression of the impulsive noise that occurs in audio recording, the noise detection method and the interpolation of audio signals have been proposed[3]. Although, these methods that have been used for recovering the corrupted musical recording is effective for only the noise that sustains within 1ms.

By the way, the signal separation methods by source properties have been proposed. In Ref. [6], a mixture of a texture image that has periodic structure and a natural image is separated by a sparse signal representation[4] with a pair of function sets(dictionaries) that are assigned to two different components. In this method, the separation is performed under the assumption that one of the signal sources can be represented by small number of the bases included in one of the dictionaries.

In this paper, we apply the signal separation method based on the sparse representation to the short duration noise suppression. For speech and noise separation, we assumed that the speech is stationary within 20-40ms and the duration of the noise is shorter than this period. In our approach, a sparse representation is employed to separate the noise and speech by difference of their time duration properties. By the sparse representation two dictionaries that support shorter or longer time period represent the noise and the speech respectively with a penalty of sparseness. In echoic environments the reverberation of the noises appears in the separated speech signals, since the duration of the reverberation of the noise exceeds the period that is supposed for the duration of the noises. To suppress the reverberation of the noise, we apply a spectrum subtraction to the separated speech. For the spectrum subtraction, we propose a power estimation method for the noise reverberation. In Sect. 2, the basic of the signal separation using a sparse representation is explained. The dictionaries for the short-duration noises are specified in Sect. 3. Next, the reverberation suppression using a spectrum subtraction is discussed. Finally, we demonstrate the suppression of the short duration noise that is recorded in an echoic environment.

2. Sparse Signal Representations and Signal Separation

Suppose we observe the signal $f$ is generated according to the model

$$f = f_s + f_n$$

where $f_s$, $f_n$ are the clean signal and the noise, respectively. We assume that two different dictionaries $A_s, A_n \in \mathbb{R}^{N \times N}$ obtains the signal decomposition

$$f_s = A_s s_x \quad \text{and} \quad f_n = A_n s_n$$

$$f = f_s + f_n$$

where $f_s$, $f_n$ are the clean signal and the noise, respectively. We assume that two different dictionaries $A_s, A_n \in \mathbb{R}^{N \times N}$ obtains the signal decomposition

$$f_s = A_s s_x \quad \text{and} \quad f_n = A_n s_n$$

By the way, the signal separation methods by source
Figure 1: Arrangement of the pair of the dictionaries
where the number of the non-zero components in \( s_x \) and \( s_n \)
is very small. Under this assumption, we estimate both of the
coefficient vector \( s_x \) and \( s_n \) from the noisy observation \( f \).
Now, the estimate vector is denoted as \( \hat{s} = [\hat{s}_x^T, \hat{s}_n^T]^T \)
where \( \hat{s}_x \) and \( \hat{s}_n \) are estimates of the clean signal and the
noise, respectively. Our only prior knowledge about the
clean signal and the noise is that the small number of the
basis vectors included in \( A_x \) and \( A_n \) can represent \( f_x \) and
\( f_n \). We hence estimate the coefficients for the clean signal
\( s_x \) and the noise \( s_n \), with a penalty that is imposed for the
sparseness of the estimates as:
\[
\hat{s} = \arg \min_{s} ||s||_p \text{ subject to } f = [A_x, A_n]s. \tag{3}
\]
The penalty function \( ||s||_p \) is Lp-norm of the vector \( s \). If
\( p = 0 \), the L0 norm indicates the number of the non-zero
components in \( s \). In this case, the signal is represented by
the minimum number of the basis vectors. However, the
L0 minimization problem in Eq.(3) is difficult to solve and
is not practical for real applications. By this reason, L0 norm penalty is replaced with L1 norm that is a convex
function. The signal decomposition problem under the L1 norm penalty is called as basis pursuit that is proposed in
Ref.[4]. In generally, Eq. (3) is reformulated as a linear program to solve. If \( A_n \) and \( A_x \) are orthonormal complete
matrices, the approximation of the solution of (3) that is the solution of
\[
\hat{s} = \arg \min_{s \in C^M} \frac{1}{2}||f - As||_2^2 + \lambda||s||_1 \tag{4}
\]
can be obtained by a fast algorithm that is proposed in
Ref.[5]. This estimate is derived under the assumption that
the Gaussian noise is added to the observation \( f \) and is referred as basis pursuit denoising (BPDN). When the dictionary
consists of orthonormal bases, Eq. (4) can be solved by a simple iterative procedure that is referred as block co-
ordinate relaxation (BCR) method[5], [6]. In this work, the
BCR method is employed to get the estimates. For the signal separation by Eq.(3), it is important to choose the
dictionary that can decompose the signal or noise component
with the smaller number of the basis vectors included in
the dictionary. In next section, we show the choice of the
dictionaries for the speech and the noise burst separation.

3. Noise Burst Suppression for Speech Signals

The duration of noises under consideration in this paper
is shorter than the period during the speech tends to
be stationary. By this assumption, we choose a dictionary
as a set of bases that support shorter time period to repre-
sent the noise components. In contrast, another dictionary
is chosen as a set of bases that support longer time period
to represent the speech components. To employ the BCR
method, we chose a set of block DFT bases that support
shorter or longer time periods. Fig. 1 shows the schematic
explanation of our method. The input speech is divided
into frames with 50% overlap. Each frame consists of \( N \)
samples. The BPDN algorithm is applied to each signal
segment divided by the frame. The pair of the dictionaries
is defined as block DFT bases. For the speech components
the separated signals. After the windowing, each separated
signal segment is added to the speech or noise outputs.

In Fig. 2, an example of the noise burst suppression
is shown. In this example, a speech signal was sampled
at 16KHz. The frame length \( N \) and the factor \( K \) were
specified as 1024 and 32, respectively. The duration of
a noise bursts was specified as 10ms. In this time interval,
the sequences of Gaussian random values are added to
the original speech signal. The top panel in Fig. 2 shows
the input signal that is a mixture of the original speech
and the noise bursts. The second and third panel shows
the separated speech and the noise signal, respectively. In
the second panel, we can see that the burst noises are suc-
cessfully removed from the original signal. In this exam-
ple, the assumption about the noise duration exactly holds.
However, in echoic environments, the reverberations of the noise burst appear at the input signal. The duration of the reverberations will exceed the frame length $N$. So, some components of the reverberations will be separated to the estimates $\hat{s}_x$ that represents the speech components. So, the leakage due to the reverberation then appears in the separated speech. Next section, we propose reverberation estimation and a suppression method based on a spectrum subtraction.

4. Noise Reverberation Suppression by Spectrum Subtraction

The reverberation of the noise is observed as a convolution between the room impulse response and the noises. In some studies, the impulse response is modeled as a white noise of which envelope decays exponentially\cite{7}. We suppose that the duration of the noise burst is so short that the distribution of the amplitude spectrum of the noise spreads in the frequency domain. From the model of the room acoustic and the frequency characteristic of noise bursts, we assume that the reverberation can be approximated as a Gaussian noise in each analysis frame. We apply the simple spectrum subtraction with a constant noise power $\sigma$ to the $k$-th component of the speech estimates $\hat{s}_x$ that is

$$
\tilde{s}_{x,k} = \begin{cases} 
\hat{s}_{x,k} & |\hat{s}_{x,k}| > \sigma \\
0 & |\hat{s}_{x,k}| \leq \sigma
\end{cases}
$$

To suppress the reverberation of the noise, one problem arises. How to specify the noise power $\sigma$? Usually, the noise power is estimated at the interval where the speech is not active. However, for noise burst, it is difficult to estimate noise power before the noise suppression. The power of the reverberation has to be estimated in a current analysis frame. In Fig. 3, measured histograms of amplitudes of the complex coefficients for 1,000,000 random variables with Gaussian PDF. The top panel in Fig. 3 shows the measured histograms of the amplitudes of the DFT coefficients. These histograms correspond to Rayleigh PDF. The bottom panel shows the histograms of the estimates $\hat{s}_{x,k}$ and $\hat{s}_{n,k}$ obtained by the BPDN. For Gaussian noises the PDF of the amplitude of the complex coefficients in two estimates are same since the Gaussian noise has no strong correlation with a particular dictionary. Comparing the histograms of BPDN and DFT, the PDF of the BPDN estimates decreases along amplitude monotonically. Since we assumed that the number of the estimated coefficients for speeches is small, so the coefficients appear as outliers of the distribution in the bottom panel of Fig. 3. In order to estimate the noise distribution due to the reverberation of the noise, we employ the median value of the coefficients in $\hat{s}_{n,k}$ to suppress the influence of the outliers coefficients due to the speech components. It is expected that the median value of $\hat{s}_{n,k}$ is proportional to the power of the reverberation of the noise, so we specify the constant $\sigma$ as $C$ times of the median value of components in a noise estimates. $C$ is a constant and is chosen as 5 experimentally. Fig. 4 shows the median of $\tilde{s}_{x}$ estimated from a mixture of a noise bursts and a speech for each analysis frame. In this experiment, the sound of hand claps that is generated in an echoic environment is chosen for the noise burst. We can see that the median varies with the power of the reverberation of the noise bursts.

5. Examples of Noise Burst Suppression in an Echoic Environment

In this section, we show some results for the speech signals corrupted by noise bursts recorded in an echoic environment. For the experiments the sound of hand claps shown in Fig. 4 was used for the noise bursts. The noise
Table 1: Results of the suppression

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input SNR(dB)</td>
<td>19.9</td>
<td>-9.4</td>
</tr>
<tr>
<td>BP</td>
<td>16.0</td>
<td>0.2</td>
</tr>
<tr>
<td>BP with SS</td>
<td>15.4</td>
<td>1.7</td>
</tr>
</tbody>
</table>

Figure 5: (a) Noisy speech, (b) separated speech and (c) noise burst is recorded in the room of which reverberation time is estimated as 0.5s. The interval of the occurrence of the noise burst was 0.2s. Speech signals were taken from ASJ-JNAS database (Japanese newspaper article sentences speech corpus). The speech signals were uttered by two male and two female speakers. Both the speech signals and the noise bursts were sampled by a rate of 16kHz. For evaluation we employed segmental SNR. Usually, segmental SNR is computed as an average of SNR computed within each frame. However, the power of the noise burst varies from minimum to maximum, the SNR of each frame of the corrupted signal varies 40dB to -20dB. The average of the SNR is not suitable for evaluation of the noise burst suppression. In order to evaluate the noise suppression, the average SNR is computed from the frames where the input SNR is lower than 6dB. For evaluation of the corruption of the speech due to the proposed method, the average SNR is computed from the frames where the input SNR is higher than 6dB. Table 1 shows the average SNR of the results of the noise suppression. We can see that the average output SNR of the frames where the input SNR is higher than 6dB decreases to about 15dB. On contrary, the SNR is improved to about 9dB for the corrupted frames. Moreover, the reverberation suppression based on the spectrum subtraction improves SNR about 1.5dB for male and female speeches. In Fig. 5 the results of the separation are shown. The top panel in Fig.5 shows the noisy speech. The second and third panel shows the separated speech and the noise signal, respectively. Some speech components that include impulsive and transient signals appear in the separated noise signal. However, the noise bursts are well suppressed in the separated speech.

6. Conclusions

In this paper, we have proposed a short duration noise suppression method based on a sparse signal representation. Two dictionaries are specified to represent the speech and the noise burst to separate from their mixtures. The separation results show the two DFT dictionaries are enough to separate the noise burst and speech signals. Moreover, under the assumption that the reverberation of the noise can be approximated as a white Gaussian noise, we propose a noise power estimation to suppress the reverberation of noises with spectrum subtraction.

References


A noise variance estimator for speech enhancement based on MAP spectral amplitude estimation

Arata Kawamura and Youji Iiguni

Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531 JAPAN
Email: kawamura@sys.es.osaka-u.ac.jp, iiguni@sys.es.osaka-u.ac.jp

Abstract—A spectral amplitude estimator for acoustical background noise suppression based on maximum a posteriori (MAP) estimation and super-Gaussian statistical modeling of the speech DFT amplitudes is previously proposed. The spectral amplitude estimator, however, cannot track to the change of the noise variance. In this paper, we introduce a new noise variance estimator into the spectral amplitude estimator based on MAP estimation. The proposed noise variance estimator which utilizes an adaptive line enhancer can estimate the noise variance whether the speech exists or not.

1. Introduction

In recent years, research on methods of speech enhancement from a speech degraded by an additive background noise is actively being done. One of the most useful speech enhancement techniques is the MMSE-STSA (Minimum Mean Square Error Short Time Spectral Amplitude) [1]. The MMSE-STSA estimates a clean speech spectral amplitude by minimizing a statistical error criterion. The MMSE-STSA can extract a speech signal with high quality and hardly produces a musical tone that is an unpleasant artificial noise caused in a spectral domain method [2]. Many speech enhancement methods based on the MMSE-STSA have been proposed [3]–[5]. These methods assume that real and imaginary parts of speech DFT coefficients are modeled as zero-mean independent Gaussian with equal variance. However, speech signals have a super-Gaussian distribution and its spectral coefficients are also a super-Gaussian distributed. To improve the speech enhancement capability, various statistical models of the speech DFT coefficients have been proposed [6]–[8]. Especially, a probability density function (PDF) proposed by T. Lotter and P. Vary is very close to that of a practical speech DFT amplitude [8]. The PDF is used in a speech enhancer based on maximum a posteriori (MAP) estimation and the extracted speech quality is improved in comparison to the MMSE-STSA. However, the speech enhancer cannot track to the change of the noise variance that is required to estimate the speech spectrum.

A noise variance estimator which adaptively track to the noise has been proposed [9]. The noise estimator selects a minimum spectral power in some past spectral powers for estimating the noise variance. However, it is difficult to tune the system parameters so that the noise variance is accurately estimated. As the other noise variance estimator, a weighted noise estimation has been proposed [10]. The weighted noise estimation is a method based on SNR (Signal to Noise Ratio), and it can detect the noise variance whether a speech exits or not. A speech enhancer with the weighted noise estimation [10] satisfies the noise suppression criterion settled by 3GPP (The Third Generation Partnership Project) [12]. The weighted noise estimator however cannot distinguish a suddenly caused noise and a speech, and the noise remains in the extracted speech.

In this paper, we propose a noise variance estimator that can detect the suddenly caused noise whose spectral amplitude keeps a large value for a long time. The proposed noise estimator utilizes ALE (Adaptive Line Enhancer) [13] which can estimate only the suddenly caused noise.

2. Conventional Method

2.1. Speech enhancement based on MAP estimation

In this section, we explain a conventional speech enhancement technique which is based on MAP estimation [8].

An observed signal $y(l)$ at time $l$ is expressed as

$$y(l) = s(l) + d(l),$$

where $s(l)$ is a speech signal and $d(l)$ is a background noise. After segmentation and windowing with a Hanning window function $h(l)$, The DFT coefficient of frame $n$ and frequency bin $k$ is represented by

$$Y(n,k) = \sum_{l=0}^{L-1} y(nQ+l)h(l)e^{-j2\pi kl/L},$$

where $L$ denotes the frame length, and the shift of the DFT is $Q$ samples. The DFT of the observed signal is represented as

$$Y(n,k) = S(n,k) + D(n,k),$$

where $S(n,k)$ and $D(n,k)$ denote the speech and the noise spectrum, respectively. The observed spectrum is also expressed as

$$R(n,k)e^{j\theta(n,k)} = A(n,k)e^{j\alpha(n,k)} + B(n,k)e^{j\beta(n,k)},$$

703
where $R(n, k), A(n, k), B(n, k)$ denote the amplitudes of $y(l), s(l), d(l)$, respectively, and $\theta(n, k), \alpha(n, k), \beta(n, k)$ denote the respective phase.

To estimate the speech spectrum, we utilize the a priori and a posteriori SNRs given by

$$
\hat{\xi}(n, k) = \frac{E[A^2(n, k)]}{\lambda(n, k)}, \quad \hat{\gamma}(n, k) = \frac{R^2(n, k)}{\lambda(n, k)},
$$

(5)

where $\hat{\xi}(n, k)$ denotes an estimated speech, $\hat{\gamma}(n, k)$ denotes the a priori SNR, $\gamma(k, n)$ denotes the a posteriori SNR, and $\lambda(n, k)$ is the noise variance. The a priori SNR is calculated as [8]

$$
\hat{\xi}(n, k) = \alpha_{\text{SNR}} \frac{A^2(n - 1, k)}{\lambda(n, k)} + (1 - \alpha_{\text{SNR}}) F[\hat{\gamma}(n, k) - 1],
$$

(6)

where $\alpha_{\text{SNR}}$ is a forgetting factor that is less than unity. From the a priori and a posteriori SNRs, we must determine a spectral gain $G(n, k)$. A speech spectrum is estimated as

$$
\hat{S}(n, k) = G(\hat{\xi}(n, k), \hat{\gamma}(n, k)) \cdot Y(n, k).
$$

(7)

The many methods to obtain the spectral gain exist. Here, we explain the spectral gain estimation method based on MAP estimation [8]. In the derivation of the speech spectral estimator, we omit $(n, k)$.

We assume that the PDF of the noise DFT coefficient is Gaussian. Then the PDF of the real part of the noise spectrum is represented as

$$
p(D_{Rk}) = \frac{1}{\sqrt{\pi \lambda}} \exp \left\{ -\frac{D_{Rk}^2}{\lambda} \right\}.
$$

(8)

The imaginary part has the same PDF. Using Eq.(8) and assuming an independence between the real and imaginary part of the DFT, the a posteriori PDF $p(Y|A)$, is expressed as

$$
p(Y|A, \alpha) = \frac{1}{\pi A} \exp \left\{ -\frac{Y - A e^{i\theta}}{A} \right\}.
$$

(9)

The PDF of the speech amplitude spectrum $A$, that is proposed in [8], is given by

$$
p(A) = \frac{\mu^{\nu + 1}}{\Gamma(\nu + 1) \sigma_S^{\nu + 1}} A^\nu \exp \left\{ -\frac{A}{\sigma_S} \right\},
$$

(10)

where $\sigma_S^2$ is a speech spectrum variance and $\nu$ and $\mu$ are the parameters that determine the PDF envelope of $A$. We assume the PDF of the speech phase $p(\alpha) = 1/2\pi$, and

$$
p(A, \alpha) = \frac{1}{2\pi} p(A).
$$

(11)

The MAP estimation values of $A$ and $\alpha$ maximize $p(A, \alpha|Y)$, then we have

$$
\hat{A} = \arg \max_A p(A, \alpha|Y) = \arg \max_A \frac{p(Y|A, \alpha)p(A, \alpha)}{p(Y)} = \arg \max_A \frac{p(Y|A, \alpha)p(A, \alpha)}{p(Y)}.
$$

(12)

$$
\hat{\alpha} = \arg \max_\alpha p(A, \alpha|Y) = \arg \max_\alpha \frac{p(Y|A, \alpha)p(A, \alpha)}{p(Y)}.
$$

(13)

The $\hat{A}$ and $\hat{\alpha}$ maximize Eqs.(12) and (13), respectively, and they also maximize the logarithms of the numerators of Eqs.(12) and (13) represented as

$$
\log(p(Y|A, \alpha)p(A, \alpha)) = \log \left( \frac{\mu^{\nu + 1}}{2\pi^3 \lambda \sigma_S^{\nu + 1}} \right) Y + \log \frac{A}{\lambda}.
$$

(14)

Differentiating Eq.(14) with respect to $\alpha$ yields

$$
\frac{\partial}{\partial \alpha} \log(p(Y|A, \alpha)p(A, \alpha)) = -\frac{(Y^* - A e^{-j\theta})(-jA e^{j\theta}) + (Y - A e^{j\theta})(jA e^{-j\theta})}{\lambda}.
$$

(15)

Setting Eq.(15) to zero, we have

$$
\hat{\alpha} = \theta.
$$

(16)

We see that the MAP estimation of the phase is simply the observed phase. We differentiate Eq.(14) with respect to $A$, we have

$$
\frac{\partial}{\partial A} \log(p(Y|A, \alpha)p(A, \alpha)) = -\frac{(Y^* - A e^{-j\theta})e^{j\theta} + (Y - A e^{j\theta})e^{-j\theta}}{\lambda} + \frac{\nu}{A} + \frac{\mu}{\sigma_S}.
$$

(17)

Setting above equation to zero, we have

$$
A^2 + \lambda \left( \frac{\mu}{2\sigma_S} - R \right) - \frac{\nu}{2} = 0.
$$

(18)

Solving the above quadratic equation, we have the spectral gain represented as

$$
G = u + \sqrt{u^2 + \frac{\nu}{2}\lambda}, \quad u = \frac{1}{2} - \frac{\mu}{4\sqrt{\lambda} \Gamma(\nu + 1)}.
$$

(19)

The enhanced speech is obtained from Eq.(7) with Eq.(19).

### 2.2. Estimation of noise variance

The spectral gain estimator (19) is easy to realize, and achieves high noise reduction capability. However, the literature [8] gives no explanation about a noise variance estimation that is required for the speech enhancement. In this section, we give a brief explanation about a noise variance estimator based on a weighted noise estimation [10].

In the $n$th frame, a posteriori SNR $\hat{\gamma}(n, k)$ is estimated by

$$
\hat{\gamma}(n, k) = \frac{|Y(n, k)|^2}{\lambda(n - 1, k)}.
$$

(20)

The weight $W(n, k)$ for $\hat{\gamma}(n, k)$ is calculated by using a non-linear function shown in Fig.1, where $\gamma_1, \gamma_2, \theta_\gamma$ are constants. An instantaneous noise power spectrum $z(n, k)$ and the noise variance $\lambda(n, k)$ are given as

$$
z(n, k) = W(n, k)|Y(n, k)|^2
$$

(21)

$$
\lambda(n, k) = \frac{\text{trace}[\mathbf{Z}(n, k)]}{\Psi(\mathbf{Z}(n, k))}.
$$

(22)
where \( \text{trace} \{ \cdot \} \) denotes the sum of elements of \( \mathbf{Z}(n,k) \), \( \Psi(\mathbf{Z}(n,k)) \) denotes the number of non-zero elements of \( \mathbf{Z}(n,k) \). We update \( \mathbf{Z}(n,k) \) as follows:

\[
\mathbf{Z}(n,k) = \begin{cases} 
\mathbf{z}(n), & n \leq T_{\text{init}} \\
\mathbf{z}(n), & \hat{\gamma}(n,k) \leq \theta_L \\
\mathbf{z}(n-1,k), & \text{otherwise} 
\end{cases} 
\]  

(23)

\[
\tilde{\mathbf{Z}}(0,k) = \mathbf{0}_{1 \times (L_z-1)} 
\]

(24)

\[
\tilde{\mathbf{Z}}(n,k) = \mathbf{Z}(n,k) \left[ \mathbf{1}_{L_z-1} \mathbf{0}_{1 \times (L_z-1)}^T \right]^T 
\]  

(25)

where \( L_z \) denotes the size for averaging the noise variance. The matrix \( \mathbf{1}_{L_z-1} \) denotes the unit matrix and its size is \( (L_z-1) \times (L_z-1) \). In the weighted noise estimation, an initial \( T_{\text{init}} \) frames are assumed that the speech signal does not exist. The noise variance \( \hat{\lambda}(n,k) \) is not updated when the a posteriori SNR is greater than \( \theta_L \) so that the noise spectrum cannot be accurately estimated.

### 3. Proposed noise variance estimator

When sinusoidal noises are suddenly caused and keep a large amplitudes for a long time, the a posteriori SNR \( \hat{\gamma}(n,k) \) shown in (23) keeps a high value. If \( \hat{\gamma}(n,k) > \theta_L \), the estimated noise variance \( \hat{\lambda}(n,k) \) is not updated. Hence, the weighted noise estimation cannot detect the suddenly caused sinusoidal noises.

In this section, we propose a new noise variance estimation method using an adaptive line enhancer (ALE) [13] that can estimate sinusoidal noises existing for a long time. The ALE estimates the present signal by linear combination of the delayed input signals. The block diagram of the ALE is shown in Fig.2, where the ADF consists of the \( M \)th order FIR adaptive filter. The output signal of the ALE at time \( l \) is represented by

\[
y_{\text{ALE}}(l) = \sum_{m=0}^{M-1} h_m(l) y(l-\Delta - m). 
\]

(26)

where \( h_m(l) \) denotes the \( m \)th filter coefficient that is updated to minimize the mean square error \( E[e^2(l)] \). The delay \( \Delta \) must be tuned so that the auto-correlation of the speech and noise signals, except for the suddenly caused sinusoidal noise, is sufficiently removed. The ADF then estimates only the sinusoidal signals that keeps the auto-correlation for the time difference \( \Delta \). Since the MAP estimation assumes an independence between the signals in the present frame and one in the other frames, we can choose the \( \Delta \) to 1 frame length. We then estimate the noise variance as \( \hat{\lambda}(n,k) + |Y_{\text{ALE}}(n,k)| \) when \( \hat{\gamma}(n,k) > \theta_L \), where \( |Y_{\text{ALE}}(n,k)| \) denotes the power spectrum of \( Y_{\text{ALE}}(l) \). The proposed speech enhancement system is shown in Fig.3, where \( \hat{s}(l) \) denotes the enhanced speech.

We performed speech enhancement simulation for a speech degraded by a tunnel noise where we added a sinusoidal noise in the interval 20000–40000 samples. In this simulation, we put the sampling frequency to 8kHz, and parameters required in the weighted noise estimation and the MAP estimation to the same values described in [8] and [10]. Table 1 shows the parameter values used in the simulation. We put the order of the ALE to \( M = 256 \). As the adaptive algorithm for the ALE, we used the NLMS
Figure 4: Speech enhancement results

(Normalized Least Mean Square) algorithm with step-size of 0.002 represented by

$$h_m(l + 1) = h_m(l) + \frac{0.002 \times y(l - \Delta - m)}{\sum_{q=0}^{M-1} y^2(l - \Delta - q)} e(l). \quad (27)$$

The result of speech enhancement is shown in Fig.4, where (a) shows the observed signal, (b) shows the speech signal extracted by the speech enhancer without the ALE, and (c) shows the speech signal extracted by the proposed system. From Fig.4(b), we see that the suddenly caused sinusoidal noise remains in the extracted speech, although the tunnel noise is successfully removed. We see that from Fig.4(c) the proposed system removes both the tunnel noise and the suddenly caused noise.

4. Conclusion

In this paper, we proposed a new noise variance estimator which can be incorporated with the spectral amplitude estimator based on MAP estimation. The proposed noise variance estimator utilizes the ALE to estimate the suddenly caused sinusoidal noises. The computer simulation gives that the proposed noise variance estimator can estimate the suddenly caused noise that keeps a large amplitude for a long time.

References


SOLITON EQUATION OF CONNECTION ENERGY
FOR SPEECH SPECTRAL INTERPOLATIONS

Yoshinao SHIRAKI

Department of Information Science, Shonan Institute of Technology,
1-1-25 Tsujido-Nishikaigan, Fujisawa, Kanagawa 251-8511, Japan
Email shiraki@info.shonan-it.ac.jp Tel: +81 466 30 0206, Fax: +81 466 34 5932

ABSTRACT

The object of this note is to generate a spectral sequence between given two speech spectra and to show that vector fields of given spectra and a kind of smoothness of connections derived from soliton equations on the field play an important role in maintaining speech intelligibility.

1. INTRODUCTION

Let us consider an interpolation function between a pair of given speech spectral sequences. It is the first difficulty in getting a smooth interpolation that speech characteristics distribute in two (time and frequency) or more higher dimensions. And the second difficulty is that speech has a discrete (or local) structure of phonemes, which limit the distribution regions of the first and the second formant frequency. In other words, the smooth interpolation considered here needs to satisfy the maintenance of the local-discrete structure as well as the smooth functions over two dimensions or more. Furthermore, such an interpolation function is not uniquely determined in general by these two conditions. In this note, we propose a method which determines the interpolation function by adding a condition of minimizing a functional of connections derived from soliton equations. Here, we do not fix the one connection. We list up many of connections and select the connection best suitable for the conditions. To do this, we use a functional of connections and investigate the functional by the variational scheme.

2. CONNECTIONS AND THEIR FUNCTIONAL

Let \( Y_A \) and \( Y_B \) be speech power-spectrum sequences of the same sentence uttered by a speaker A and B, respectively. Here, power-spectra are represented by Autoregressive (AR) model[5, 7, 11]. Let us examine the whole map from \( Y_B \) to \( Y_A \); \( H := \{ \phi : Y_B \rightarrow Y_A \} \). In order to combine the theory here to practical algorithm, we use two-step examinations of the map \( F \). Note that \( Y_B, Y_A \) are curved surfaces. First, \( Y_B, Y_A \) are projected respectively by a map \( \phi \). Second, we investigate the map \( h^\phi : Y_B^\phi \rightarrow Y_A^\phi \) between projected \( Y_B^\phi, Y_A^\phi \).

2.1. Connection, smoothness and linearity

A connection \( \nabla \) is a kind of a measure for the smoothness of a given surface \( Y \). For example, the Euclidean space \( \mathbb{R}^n \) has a flat connection and the sphere \( S^2 \) has a non-flat connection. A connection \( \nabla \) also measures the smoothness along with a given line \( L \) on the surface \( Y \). This smoothness of the line \( L \) locally depends on a vector field \( X \) on the surface \( Y \). In other words, a connection is a function of two tangent vector fields \( X \) and \( Z \), where \( X \) is a tangent vector field along with the line \( L \) and \( Z \) is a tangent vector field on the surface \( Y \). This fact that a connection is a measure for the smoothness of a higher dimension space is the reason why we adopt a connection energy as the additional smoothness condition for the map \( h \). For two vector fields \( X, Z \) and a function \( f \), a connection \( \nabla \) satisfies the following equations:

\[
\nabla fXZ = f\nabla XZ, \quad \nabla fZ = f\nabla XZ + X(f)Z
\]

These two equations mean that a connection \( \nabla \) is a derivative with respect to \( X \) of \( Z \), and a connection is linear for \( X \), but is not always linear for \( Z \). Note that we can represent a connection \( \nabla \) in the local coordinates \( \{ U_j \} \). So, we abbreviate the upper equation to \( \nabla = \partial + A_\alpha \) using \( k \times k \) matrices differential 1-form \( A_\alpha \). Now, for \( \nabla = \partial + A_\alpha \) on \( U_\alpha \), the curvature is defined by \( F_\nabla = dA_\alpha + A_\alpha \wedge A_\alpha \).

2.2. Connection energy and nonlinear harmonic form

Let \( A \) be a connection. A connection energy functional is defined by the following equation using the curvature \( F_A \):

\[
Ym(A) := \int_Y |F_A|^2 * 1_Y
\]  

(1)

Here, \( *1_Y \) is a volume form of \( Y \). This energy function is called the Yang-Mills functional and is an essential notion for examining the smoothness structure of manifolds \[1, 2, 3\]. It is well known that the Yang-Mills functional has some soliton equations as its special solutions\[4\]. Let us
define an exterior derivative \( d_A \) associated with connection \( A \) and also define \( d_A^* \) as the adjoint operator of \( d_A \). Using these operators, we can define nonlinear Laplacian operator as \( \Delta_A := d_A d_A^* + d_A^* d_A \). Using this nonlinear Laplacian operator \( \Delta_A \), for the curvature \( F_A \) minimizing equation (1), the following equation holds:
\[
\Delta_A F_A = 0.
\]
This equation means that the curvature \( F_A \) is a nonlinear harmonic form. For simplicity, let us consider the case of the line bundle. Now, let \( \delta \) be a usual exterior derivative and \( \delta \) be its adjoint in the sense of \( L^2 \) inner products. Then, for usual Laplacian \( \Delta := \delta \delta^* + \delta^* \delta \), the following equation holds:
\[
\Delta F_A = 0. \tag{2}
\]
Note that, in 4-dimensional case, the Laplacian operator \( \Delta_A \) is deeply concerned with the boundary operator \( \partial \) introduced in [10].

2.3. Flat connection and velocity of spectral change

The flat connection is one of the connections satisfying equation (2). Therefore, a piece-wise linear space has a local flat connection in the linear space while the whole connection is not flat. Now, if the operator \( \Delta \) in equation (2) is \( \frac{\partial^2}{\partial t^2} \), for constants \( C_0, C_1 \), one of solutions satisfying equation (2) is represented as the following:
\[
F_A(t) = C_0 t + C_1, \quad t \in [t_0^P, t_2^P] \tag{3}
\]
This equation means that any functions satisfying equation (2) have a linear form as in equation (3) with respect to time \( t \). On the other hand, we define a velocity function \( \Theta(t) \) of spectral change at time \( t \) by the following equation:
\[
\Theta(t) = ||f(t + dt) - f(t)||^2 \tag{4}
\]
Here, \( f(t) \) is the power-spectrum at time \( t \) and the norm \( || \cdot ||^2 \) is the squared integral of a log-scale power spectrum in the interval \([-\pi, \pi] \). In general, as we can assume the power spectrum \( f(t) \) changes slowly in a short time interval \([t_0^P, t_2^P] \), for constants \( D_0, D_1, \Theta(t) \) satisfies the following equation:
\[
\Theta(t) = D_0 t + D_1, \quad t \in [t_0^P, t_2^P] \tag{5}
\]
where an interval \([t_0^P, t_2^P] \) is one of local coordinates of a piece-wise linear space. This equation means that because the velocity function \( \Theta(t) \) has a property of local linearity, the function \( \Theta(t) \) satisfies equation (2) in the local interval \( t \in [t_0^P, t_2^P] \). Furthermore, for two speech spectra \( Y_A \) and \( Y_B \), by setting \( C_1 = D_1 = 0 \), their velocity functions \( \Theta_A(t) \) and \( \Theta_B(t) \) have a relationship in the common time interval \([t_0^P, t_2^P] \) as the following equation:
\[
\Theta_A(t) = \alpha_0 \Theta_B(t), \quad t \in [t_0^P, t_2^P] \tag{6}
\]
Therefore the velocity function of \( Y_B \) is locally equal to that of \( Y_A \) except a constant factor \( \alpha_0 \).

3. Algorithm : Linear Interpolation

We give algorithms for determining the connection satisfying equation (2). In fact, for the following approximation based on the local linear interpolation, the connection satisfies equation (2) locally with respect to time \( t \). So, we can generate a kind of approximation of \( Y_A \), and get the map \( \varphi : Y_A \to Y_A^{\varphi} \). By using this approximation, a projected spectrum \( Y_A^{\varphi} \) satisfies (2). Now, let us consider the practical problem of how to deform \( Y_B \) gradually to \( Y_A \) while maintaining intelligible speech. In this case, we have to find the smooth map both in time and frequency domain.

3.1. Interpolation in time domain

In time domain, before we use a linear interpolation technique developed in [?], we use Dynamic Time Warping (DTW) as the time alignment function from \( Y_B \) to \( Y_A \). The following is the procedure. Let \( T_{BA} \) be the DTW operator and \( T_{BA}(Y_B) \) be the spectral sequences after the operation \( T_{BA} \), which minimizes the sum of spectral distortions of \( Y_A \) and \( Y_B \).

The followings are the linear interpolation procedures in time domain. First for each interval \( t \in [t_{BA}^{\varphi^{-1}}, t_{BA}^{\varphi}] \) we approximate \( Y_A \) by interpolating two spectra based on the least square method as the following equation:
\[
y_A^{\varphi}(t) = y_A(t_{BA}^{\varphi}) \phi_A^{\varphi}(t) + y_A(t_{BA}^{\varphi^{-1}}) \phi_A^{\varphi^{-1}}(t) \tag{7}
\]
Here, \( y(t) \) is the power-spectrum of \( Y \) at time \( t \) and time points \( t_{BA}^{\varphi}(i = 1, \cdots, m) \) are determined so that the sum of spectral distortion between original spectrum \( Y_A \) and the linearly approximated spectrum is minimized [?]. We call \( \phi_A^{\varphi}(t) \), \( \phi_A^{\varphi^{-1}}(t) \) time function of \( Y_A \).

Next we interpolate \( y_B(t) \) linearly by time function \( \phi_A^{\varphi}(t) \) and \( \phi_A^{\varphi^{-1}}(t) \) of a speaker \( A \) as the following equation:
\[
y_B^{\varphi}(t) = y_B(t_{BA}^{\varphi}) \phi_A^{\varphi}(t) + y_B(t_{BA}^{\varphi^{-1}}) \phi_A^{\varphi^{-1}}(t) \tag{8}
\]
By using the property of a local linearity, we get two surfaces \( Y_A^{\varphi} \) and \( Y_B^{\varphi} \) projected by the map \( \varphi \). Because \( Y_A^{\varphi} \) and \( Y_B^{\varphi} \) are both piece-wise linear surface, curvatures of \( Y_A^{\varphi} \) and \( Y_B^{\varphi} \) locally satisfy equation (2) with respect to time \( t \).

Next, for almost every time point \( t \in [t_{BA}^{\varphi^{-1}}, t_{BA}^{\varphi}] \) we can assume the following equation holds : \( \phi_A^{\varphi}(t) + \phi_A^{\varphi^{-1}}(t) = 1 \). Therefore, we obtain the following equation:
\[
\Theta_A(t) = (\phi_A^{\varphi}(t + dt) - \phi_A^{\varphi^{-1}}(t)) ||y_A(t_{BA}^{\varphi}) - y_A(t_{BA}^{\varphi^{-1}})||^2 \tag{9}
\]
The approximation error in this equation depends on the approximation precision in equation (7). We can make the approximation precision small enough by setting \( m \) a large number. For two time-points \( t_{BA}^{\varphi^{-1}}, t_{BA}^{\varphi} \) we denote corresponding power spectrum of \( T_{BA}(Y_B) \) by \( y_B(t_{BA}^{\varphi}), y_B(t_{BA}^{\varphi^{-1}}) \), respectively.
We can represent the dynamic measure $\Theta_{B_{\alpha}}(t)$ of $Y_{B_{\alpha}}$ by the following equation:

$$\Theta_{B_{\alpha}}(t) = \phi_B(t + dt) - \phi_B(t)^2 ||y_B(t_{\alpha}^1) - y_B(t_{\alpha}^2)||^2$$

(10)

For the interval $[t_{\alpha}^1, t_{\alpha}^2]$ let set a constant $\alpha_1$ be $||y_A(t_{\alpha}^1) - y_A(t_{\alpha}^2)||^2 / ||y_B(t_{\alpha}^1) - y_B(t_{\alpha}^2)||^2$, then we obtain the following equation.

$$\Theta_A(t) \approx \alpha_1 \Theta_{B_{\alpha}}(t), \quad t \in [t_{\alpha}^1, t_{\alpha}^2]$$

(11)

Now, let us regard the velocity function $\Theta(t)$ as a kind of the dynamic measure[9]. The dynamic feature is an important characteristic of intelligibility and individuality. As mentioned before about the velocity functions, the dynamic measure of $Y_{B_{\alpha}}$ is equal to that of $Y_A$ (or $Y_A^{\alpha}$) except a constant factor $\alpha_1$. Therefore, we can practically reflect the dynamic feature of $Y_A$ to that of $Y_B$ based on the interpolation equation (8).

### 3.2. Interpolation in frequency domain

Next, we generate a map $h^\alpha : Y_{B_{\alpha}} \rightarrow Y_A^{\alpha}$. We denote time-corresponded power spectrum in log-scale by $f_A(\omega)$ and $f_B(\omega)$. Using $f_A(\omega)$ and $f_B(\omega)$, we generate a spectra $f_C(\omega)$ by the following interpolation procedure.

1. Determine a non-linear mapping in frequency domain between $f_A(\omega)$ and $f_B(\omega)$ by DFW (Dynamic Frequency Warping) scheme minimizing the LSD measure.
2. Interpolate the spectrum between frequency assigned spectra $f_A(\omega)$ and $f_B(\omega)$.
3. Generate the spectrum on the trajectory through the interpolated spectra above using IFIS[6].

Here, LSD (Laplacian Spectral Distortion) is defined by the following equation:

$$\text{LSD} = \left( \frac{1}{\omega_{\Phi}} \int_{-\pi}^{\pi} (\Delta_\omega f_A(\omega) - \Delta_\omega f_B(\omega))^2 \frac{d\omega}{2\pi} \right)^{1/2}$$

(12)

where $\Delta_\omega = \frac{\omega_{\Phi} - \omega_{\Phi}}{2\pi}$. As LSD makes much of both local peaks and bottoms of speech spectra, it is expected that the interpolated spectrum based on LSD measure maintains speech intelligibility.

### 4. EXPERIMENTAL RESULTS

Several experiments were conducted to examine the performance of the proposed interpolation method in handling speech data. The 12th order AR parameters were calculated using ordinary linear prediction analysis for a frame of 10 ms with a 30 ms-long Hamming window and 8 kHz sampling rate. The number of FFT points for a power spectrum calculation is 256. 1024 points FFT was used for the IFIS calculation of inverse function to maintain the precision. We use the utterance ‘ KoNi ni chi wa ’ spoken by a male and a female speakers. In this case, average cepstrum distance after DTW is 7.2 db.

Fig.1 shows extraction examples of the dynamic measure after DTW. In this figure, the horizontal axis represents time (×10 ms) and the upper plot is of a female speaker, a middle plot is of a male speaker modified by a female speaker, and the lower is of a male speaker. Figure 1 indicates that the dynamic measure modified by a female speaker is closer to the female’s than the male’s.

Next, we compared four interpolation methods in frequency domain; (a) Interpolation in LSP parameters (b) Interpolation by IFIS (c) Interpolation by DFW(MF) + IFIS (d) Interpolation by DFW(FM) + IFIS. Here, DFW(MF) represents the interpolation from a male to a female speaker and DFW(FM) represents the interpolation from a female to a male speaker. Fig.2 shows interpolated spectra. The lowest spectrum is of a male speaker. This figure shows that a proposed method (d) is superior to others from the standpoint of making much of maintaining peaks and bottoms of interpolated spectrum.

Next, a subjective listening test was conducted to evaluate the speech intelligibility as a specific individual of the interpolated spectrum. Eight listeners (3 females and 5 males) subjectively evaluated 5-level score of speech intelligibility as a specific individual. Four speech were presented; (1) LPC speech of a male speaker (2) LPC speech of a female speaker (3) interpolated speech based on LSP parameters (4) interpolated speech based on a proposed method (d). Interpolation ratios were 0.3, 0.4, and 0.5 from a male to a female speaker. Both of the speech power and the pitch period were also interpolated in these ratios. The number of trials was 5. Table 1 shows average scores. This table indicates that subjective quality of a proposed method is superior to that of LSP’s.

### 5. CONCLUSION

We generated a spectral sequence between given two speech spectra $Y_A$ and $Y_B$ and showed that vector fields of given spectra and a kind of smoothness of connections on the field play an important role in maintaining speech intelligibility. It has been shown from subjective listening tests that the speech intelligibility of a proposed interpolation method is superior to the conventional method.

### References


Fig. 1: Plots of dynamic-measure after DTW (by a female, by a male modified by a female, by a male).

Fig. 2: Some examples of spectral interpolation between a male and a female speaker.

Table 1: Subjective listening results

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
<th>LSP</th>
<th>Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.80</td>
<td>4.60</td>
<td>1.81</td>
<td>2.15</td>
</tr>
</tbody>
</table>
Unsupervised structuring element optimization of morphological opening for texture images

Atsusi Fujiki†, Akira Asano‡, and Mitsuji Muneyasu†

†Faculty of Engineering, Kansai University
Yamatecho 3-3-35, Suita, Osaka 564-8680, JAPAN
‡Graduate School of Engineering, Hiroshima University
Kagamiyama 1-7-1, Higashi-Hiroshima, Hiroshima 739-8521, JAPAN
Email: asano@mis.hiroshima-u.ac.jp

Abstract—A method of optimizing structuring elements of morphological openings for extracting structures of texture image is proposed. This method can adopt the structuring element to the microstructure of a texture image, even if it is corrupted by noise. This method is based on the property of the texture that it is composed of a repetitive appearance of the microstructure. The extraction ability from noisy images is improved by introducing limitations to the variance of the pixel values within the structuring element in the process of optimization, and iterating the optimization with modifying the size of the structuring element. As an example of the application, the effective noise removal is achieved by the opening using the estimated shape of the microstructure as the optimal structuring element.

1. Introduction

Opening is one of the most important image operations in the context of mathematical morphology [1–3]. Opening presents the composition of an image by the repetitive arrangement of a structuring element, which is a small object used as a probe. The significance of opening is its quantitativeness in the sizes of image objects. For example, a quantitative noise removal in images is achieved by opening in the sense that noise objects smaller than the structuring element are removed exactly.

Since the shape of the structuring element appears directly in the result of opening, the opening using the structuring element resembling objects contained in the target image preserves the visual appearance of the whole image. Of course it is not generally possible to determine one typical object resembling various objects contained in an image. However, if the target image is restricted to a texture, we can derive a typical object representing the whole texture.

The shape of repetitively appearing objects in a texture should be preserved by the optimal opening for texture characterization, since this is the fundamental characteristic for representing the visual appearance of the target texture. In the case that opening is applied for noise removal, a structuring element not resembling the repetitively appearing objects causes in the output image undesired microstructures which are not related to the original image.

The optimal opening described above is achieved by using this typical object as the structuring element.

Such structuring element can be estimated by using “Primitive, Grain, and Point Configuration (PGPC)” texture model [4] which we have proposed as a model of texture description. The PGPC texture model regards a texture as an image composed by a regular or irregular arrangement of grains that are much smaller than the size of image and resemble each other, and presents a method of estimating the fundamental object, called primitive, from which the grains are derived by a certain modification.

We propose in this paper a novel optimization method which can estimate the primitive even if the target texture is corrupted by noise. It utilizes the a priori knowledge that the extent of noisy pixels is smaller than the original microstructure of textures, and introduces limitations on the shape of structuring elements based on this knowledge in the process of optimization.

We show an experimental result of an application of the proposed method for noise removal from texture images. In our previous work [5], we proposed an optimization method of gray scale opening for noise removal of texture images using the primitive estimation based on the PGPC texture model. The previous method, however, requires an example of the noise-free image of the target corrupted texture. It does not need to be the exact noise-free version of the target corrupted texture because of the characteristics of the textures described above, but should be at least a different uncorrupted realization of the texture which has the same microstructure as the target corrupted image.

The estimation method proposed in this paper can optimize the structuring element for noise removal without any noise-free example, since the proposed method can estimate the texture primitive from a noisy texture image. The experiment shows that the proposed method is as effective as the method using the noise-free example.
2. PGPC texture model and basic primitive estimation procedure

2.1. Morphological size distribution

Opening of image $X$ with respect to structuring element $B$ means residue of $X$ obtained by removing smaller structures than $B$. It indicates that opening works as a filter to distinguish object structures by their sizes. Let $2B, 3B, \ldots$ be homothetic magnifications of the basic structuring element $B$. We then perform opening of $X$ with respect to the homothetic structuring elements, and obtain the image sequence $X_B, X_{2B}, X_{3B}, \ldots$. In this sequence, $X_B$ is obtained by removing the regions smaller than $B$, $X_{2B}$ is obtained by removing the regions smaller than $2B$, $X_{3B}$ is obtained by removing the regions smaller than $3B$, \ldots. If $B$ is convex, it holds that $X \subseteq X_B \subseteq X_{2B} \subseteq X_{3B} \subseteq \ldots$. This sequence of opening is called granulometry [3].

We then calculate the ratio of the area (for binary case) or the sum of pixel values (for gray scale case) of $X_B$ to that of the original $X$ at each $r$. The area of an image is defined by the area occupied by an image object, i.e., the number of pixels composing an image object in the case of discrete images. The function from a size $r$ to the corresponding ratio is monotonically decreasing, and unity when the size is zero. This function is called size distribution function. The size distribution function of size $r$ indicates the area ratio of the regions whose sizes are greater than or equal to $r$.

2.2. PGPC texture model

The PGPC texture model regards a texture as an image composed by a regular or irregular arrangement of objects that are much smaller than the size of image and resemble each other. The objects arranged in a texture are called grains, and the grains are regarded to be derived from one or a few typical objects called primitives. This model is based on the observation, suggested by Gestalt psychology, that a repetitive appearance of similar objects of a moderate size is organized to be a meaningful structure by the human cognitive process.

We assume here that the grains are derived from one primitive by homothetic magnification. We also assume that the primitive is expressed by a structuring element $B$, and let $X$ be the target texture image. In this case, $X_B$ is regarded as the texture image composed by the arrangement of $rB$ only. It follows that $rB - (r + 1)B$ indicates the region included in the arrangement of $rB$ but not included in that of $(r + 1)B$. Consequently, $X_{3B} = X_{2B} - X_B$ is the region where $r$-size grains are arranged if $X$ is expressed by employing an arrangement of grains which are preferably large magnifications of the primitive. The sequence $X = X_B, X_B - X_{2B}, \ldots, X_{rB} - X_{(r+1)B}, \ldots$ is the decomposition of the target texture to the arrangement of the grains of each size.

2.3. Basic primitive estimation procedure

Since the sequence can be derived by using any structuring element, it is necessary to estimate the appropriate primitive that is a really typical representative of the grains. We employ an idea that the structuring element yielding the simplest grain arrangement is the best estimate of the primitive, similarly to the minimum description length (MDL) principle [6]. The simple arrangement locates a few number of large magnifications for the expression of a large part of the texture image, contrarily to the arrangement of a large number of small-size magnifications. We derive the estimate by finding the structuring element minimizing the integral of $1 - F(r)$, where $F(r)$ is the size distribution function with respect to size $r$. The function $1 - F(r)$ is 0 for $r = 0$ and monotonically increasing, and 1 for the maximum size required to compose the texture by the magnification of this size. Consequently, if the integral of $1 - F(r)$ is minimized, the sizes of employed magnifications concentrate to relatively large sizes, and the structuring element in this case expresses the texture using the largest possible magnifications. We regard this structuring element as the estimate of primitive.

We estimate the gray scale structuring element in two steps: the shape of structuring element is estimated by the above method in the first step, and the gray scale value at each pixel in the primitive estimated in the first step is then estimated. However, if the above method is applied to the gray scale estimation, the estimate often has a small number of high-value pixel and other pixels whose values are almost zero. This is because the umbra of any object can be composed by arranging one-pixel structuring element. This is absolutely not a desired estimate. Thus we minimize $1 - F(1)$, i.e., the residual area of $X_B$ instead of the above method. Since the residual region cannot be composed of even the smallest magnification, the composition by this structuring element and its magnification is the most admissible when the residual area is the minimum.

The exploration of the structuring element can be performed by the simulated annealing, which iterates a modification of the structuring element and find the best estimate minimizing the evaluation function described in the above [4].

3. Primitive estimation from corrupted image

The primitive estimation method described in the previous section has been developed for noiseless texture images. If it is applied for texture images corrupted with noise, it does not work well, since noisy pixels are incorporated to the estimation and the estimated primitive includes undesired noisy pixels.

To avoid this problem and achieve the estimation from corrupted image, we introduce the following two modification into the estimation procedure: 1) Limitation of the variance within the structuring element, and 2) Iterative
estimation with reducing the extent of the structuring element.

3.1. Limitation of the variance within the structuring element

The extent of noisy pixels is usually small, and the pixel values are often significantly different from the neighborhood pixel values of the noiseless image. Thus the primitive estimated from noisy images often has a pixel whose value is significantly and unnaturally different from the other pixels. To avoid the problem, the variance of the pixel values within the structuring element is limited in the optimization procedure.

A structuring element is generated at each iteration by the modification to the structuring element at the previous iteration. If the variance of the pixel values is larger than the threshold, this structuring element is discarded and another structuring element is newly generated. The threshold is relatively large at the beginning of the procedure to allow a large variance, and decreases along the progress of simulated annealing to make the structuring element converge to the appropriate one.

3.2. Iterative estimation with reducing the extent of the structuring element

The extent of the estimated primitive is important when the primitive is used as the structuring element for opening. The smaller the extent is, the higher the ability of preserving the details of the image is, but the lower the ability of noise removal is. Thus we introduce an iteration of the estimation with reducing the extent of the primitive, to find the optimal extent.

After the initial estimation, the primitive is estimated again using the opened noisy image by the previously estimated primitive and reducing the extent of the primitive and the limitation of the variance. These estimations are iterated to obtain the final estimate.

4. Experiments

4.1. Parameter settings

We use 8-bit gray scale texture images of $64 \times 64$ for the experiments. Example texture images are shown in Fig. 1.

The estimation by simulated annealing has a parameter called temperature that controls the probability where the modification of structuring element is accepted even if the evaluation function is increased by this modification. The temperature at the $i$th iteration in one optimization process, denoted $T_i$, is defined so that acceptance probability $P(\Delta F)$ is as follows:

$$P(\Delta F) = \frac{1}{1 + \exp(\frac{\Delta F}{T_i})},$$

where $\Delta F$ is the increment of the evaluation function. The initial temperature $T_0$ is determined by setting the acceptance probability $P$ at the initial state to 0.35 and the following calculation:

$$T_0 = \frac{\Delta F}{\log\left(\frac{1}{P} - 1\right)}.$$  \hspace{1cm} (2)

The temperature decreases following the iteration, as follows:

$$T_{i+1} = 0.98T_i.$$  \hspace{1cm} (3)

The number of iteration is fixed to 1000 times in our experiments.

The initial structuring element is set to the cross-shaped one of 9 pixels, and the initial pixel values are set to 50 at all the pixels. The number of pixels is fixed to 9 at the initial estimation, and the estimation procedures are repeated with decreasing the number of pixels by one, and the estimate where the number of pixels is 5 is regarded as the final estimate of the primitive. The final number of pixels is set to 5 based on a preliminary experiment. The threshold used for the limitation of variance is set to 500 when the number of pixels is 9, and decreases by 100 as the number of pixel decreases by one.

4.2. Results of optimization

We show here experimental results for noisy images corrupted by impulsive noise of positive values to present the effectiveness of the optimized opening. The pair of Fig. 1 (a) and (b) is extractions of different parts from the same texture. Corrupted images are generated by selecting 1000 pixels randomly in Fig. 1 (a) and adding random positive values to the selected pixels. If the noisy pixel value is larger than 255, it is replaced with 255.

Figure 2 shows the result of noise removal from the corrupted image generated from Fig. 1 (a). The estimation was tried five times, and the best results are shown here. The results by the opening optimized by our previous method [5] using noncorrupted images, and median filter, switching median filter, PSM filter [7] as typical image filter for impulsive noise, are shown for comparison. Figure 1(b) is used for the estimation by our previous method. The thresholds for the switching median filter and PSM filter are determined to yield the best results by repetitive experiments.

The mean-square-error in comparison with the original image (a) is shown at each image in Fig. 2. These results
show that the proposed method has achieved the equivalent ability of the structure extraction and noise removal to our previous method without noncorrupted example images. They also show that our optimized opening has higher ability than the typical noise removing filters. It suggests that these typical filters degrade microstructures in texture images.

5. Conclusions

This paper has proposed the optimization method of morphological opening for extracting textural structures from noisy texture images. This method achieves the optimization with the target noisy image only and without any noncorrupted example by introducing an appropriate restriction. It is shown by an experiment of impulsive noise removal in texture images that the method is more effective than typical noise removing filters.

The purpose of our method is currently limited to the extraction of structures in texture images. The fundamental characteristic of the texture used in this method is, however, that the image is composed by the repetitive appearance of similar grains derived from a primitive, which is described by the PGPC texture model. It indicates that the method can be applied to general images other than textures, if the images can be transformed to those having the property. We are now working on this.

References


A Fast Binary Morphological Operator for Real Time Sequence Analysis

Alessandro Bevilacqua and Alessandro Gherardi

ARCES (Center of Excellence on Electronic Systems)
University of Bologna, Via Toffano 2/2, 40125 Bologna, ITALY
Email: {abevilacqua, agherardi}@arces.unibo.it

Abstract—The Structural Fitness (SF) method represents a binary morphological algorithm which has been proved to be particularly effective to segment objects in very noisy contexts. However, the original algorithm is quite complex and very time consuming, thus making this method not suitable for real time sequence analyses. In this work we present an optimized algorithm which exploits the potential of a 32-bit computer architecture and the overlap between subsequent processing areas. While the original SF algorithm’s computational complexity is $O(Nn^2)$ (where $N$ is the number of image pixels and $n$ is the size of the filter mask) the optimized version’s is $O(\sqrt{Nn})$. The improvement enables the method to be used in real time applications.

1. Introduction

One of the earliest stage performed by a sequence analysis-based application is image segmentation. Most of the methods used to perform this step are based on Mathematical Morphology (MM). In particular, binary morphology operators are widespread used in a variety of sequence analysis based systems. The main advantage of using the powerful nonlinear operators of MM is that they achieve fast and localized decisions suitable to shape and texture identification. In addition, many algorithms are known which optimize time performance by exploiting the peculiarities of the 32-bit or 64-bit computer architectures. This is why they are widely used in real time sequence analysis.

However, classical binary morphology operators suffer from low precision when dealing with very noisy binary images. The Structural Fitness (SF) operator introduced in [2] and better explained in [1] has shown better qualitative results. In particular, it results to be very effective to segment objects even in the presence of structured noise. The strength of this operator is that it can look for structures which the objects are made of very accurately, thus being able to separate objects from structured noise. The drawback of the original SF algorithm arises from the complexity of the method: it is not suitable to be employed in real time sequence analyses, being very time consuming.

In this work we present an optimized algorithm which exploits the bitmap representation of binary images as well as the overlap between subsequent processing areas during the convolution process. In particular, all the logical ANDs operations required by the SF method are computed in advance (at the beginning at the convolution process) and stored onto a matrix in an opportune manner. The typical redundancy originated from a convolution process can be exploited successfully thus reducing the overall number of the operations accomplished. While the original SF algorithm’s computational complexity is $O(Nn^2)$ (where $N$ is the number of image pixels and $n$ is the size of the cell-based Structuring Element) the optimized version reduces the computational complexity to $O(\sqrt{Nn})$. The optimized algorithm becomes thus suitable to be utilized efficiently in real time sequence analysis (e.g. for visual surveillance or traffic monitoring applications).

2. Previous Works

Classic MM is widely spread to face the problem of object segmentation, due to the simplicity of use of the two basic MM operators: erode and dilate. Therefore, these have been the first operators to be optimized. One of the earliest optimization of binary erosion and dilation is due to the work described in [4]. Their method is based on SIMD (Single Instruction Multiple Data) processing. The main assumption which stands behind the method is that images must be provided in true binary form, that is 1 bpp (bit per pixel) in depth. By storing 16 adjacent pixels in a 16 bit register it is possible to process all the source image pixels in parallel through logical bitwise operators and shifting. Although the algorithm is quite simple, it works just for specific Structuring Elements (SE’s). The “destination word accumulation” (dwa) technique developed by the author in [3] extends the method developed in [4] to make it able to work with any SE. The algorithm has been conceived to write on the destination image a whole word at a time and for this reason dwa is amenable of efficient implementation. Author in [6] has shown that the shape of the SEs can induce redundancy and for this reason he uses a logarithmic decomposition method to achieve a faster computation. Although the dwa method is the fastest one it is suitable just for binary erode and dilate operators. Authors in [7] make use of critical points between SE translation to speed up computation and to compute just the values needed to each translation. This method is suitable for arbitrary SEs. Authors in [5] exploit the overlapping area while performing a raster morphological operation: overlapped pixels are computed only once.
3. Structural Fitness Method: an outline

The SF method we developed has been widely described in [1]. It has been conceived to work well even in the presence of very noisy images, such as those coming from sequences taken in outdoor environments. Therefore, our target applications regard both outdoor visual surveillance and traffic monitoring. For every moving object detected through the background subtraction technique, there are a lot of false signals which alters the right perception of the blob itself (Fig. 1). False signals may come from different sources, such as random noise, small movements in the background scene (e.g. moving trees), camera displacements due to wind load.

One of the most used methods to remove noise consists into applying, one or more times, MM operators to the binary image. This approach reveals three drawbacks. First, it is quite difficult to find out the right value for the kernel size and the operations’ sequence. Second, small foreground objects may be eliminated. The third and most important disadvantage is that morphological operations may often result into object’s shape and border distortion.

Our method aims to give a measure of how much a pixel belongs to a structural windowed region around it. To this purpose, we introduce the concept of fitness of the pixel at the center of the SE in respect of the pattern it should belong to. The first step is to define the basic structure we intend to address. Fig. 2(a) and (b) show the basic and the compound structure we use. The latter is obtained by rotating the former by 90°, 180° and 270°. This is as to say that the basic structure is searched by considering every spatial arrangement. In addition to these two structures, we define a cell-based structure (Fig. 2(c)). It is built through stemming from the compound structure (b) the same as (b) has been built starting from (a). Accordingly, the white circle are the centers of the compound (cell) elements.

Actually, all the pixels of the elements involved in (a), (b) or (c) are assigned “1”. In case of the basic structure (Fig. 2(a)), a logical AND between the pixel pointed by the circle and each one of its three neighborhoods is performed. The arithmetic sum of these three partial results represents the partial fitness of the pixel pointed by the circle (therefore, the partial fitness maximum value is 3).

In case of the compound structure, this procedure is accomplished four times and the partial fitnesses computed for the pixels pointed by the white circles are summed to each other and given to the central pixel. The same procedure is accomplished for the cell-based SE and the central pixel retains the overall fitness. If this value is lower than a given threshold $T_F$ the underline image pixel is given “0” anyway. Otherwise, the original value is preserved. Similarly to what happens in binary MM, the threshold value $T_F$ depends on the size and the shape of the SE’s, on the amount of noise of the input image (e.g., Fig. 1) and mainly on the structure of the objects we are looking for. Even though some statistical method has been devised to determine this value automatically, dealing with this topic is beyond the purpose of this work.

4. Structural Fitness Optimization

All the methods used to optimize binary morphological operations exploit the bitmapped representation of the image and of the SE as well (see Sect. 2). This is possible because performing a classical erode or dilate operation simply means performing a convolution without any further computation on partial outcome (e.g. thresholding operations). On the other hand, in the SF method is crucial to compare the fitness for a pixel with a reference threshold value, which is usually greater than one and cannot be stored in one bit. In addition, the redundancy of the original algorithm arises from overlapping regions which are “hidden” due to the way the compound and the cell-based SE’s are computed. Therefore, we rearrange the original algorithm to perform it in two stages: in the first stage all the logical ANDs necessary to the second stage to compute the very fitness value are calculated once and for all.

4.1. Stage 1: Preprocessing

The basic assumption which originates this step is that in the algorithm outlined in Sect. 3 all the logical ANDs computed in the compound SE (Fig. 2(b)) refer to two directions: horizontal and vertical. Fig. 3(a) shows the compound SE where the oval shapes represent logical ANDs. Fig. 3(b) and (c) show how these operations can be computed separately along the horizontal and the vertical direction, respectively. Since in the original algorithm same logical ANDs are computed repeatedly when calculating the overall fitness value in the cell-based SE (Fig. 2(c)), it could be useful to compute all the logical operations in advance. To this purpose, all the horizontal and vertical ANDs are arranged in a structure called Structuring AND.
Figure 3: The compound SE (a) and its components along horizontal (b) and vertical (c) directions, where “ovals” represent logical ANDs.

Figure 4: The SAND bits along the horizontal direction and each \( \bar{v}_i \) represents all the logical ANDs between adjacent pixels along the horizontal direction and each \( \bar{h}_i \) row represents the logical ANDs between pixels along the vertical direction. As an example, the shadowed cells of Fig. 4 represent the compound SE of Fig. 3(a).

Now, let \( H_{i,j} \) and \( V_{i,j} \) be the elements of the rows \( \bar{h}_i \) and \( \bar{v}_i \), respectively. Then:

\[
H_{i,j} = I_{i,j} \bigwedge I_{i,j+1}
\]

(1)

\[
V_{i,j} = I_{i,j} \bigwedge I_{i+1,j}
\]

(2)

where \( I_{i,j} \in \{0, 1\} \) represents a generic pixel of the input image (e.g. the circled cell in Fig. 4 is \( H_{2,1} \)). At last, we can use the dwa technique to compute all the rows of the SAND (\( \bar{h}_i \) and \( \bar{v}_i \)) once and for all. What follows is the simple instruction in C language to implement Eq. 1 using this technique:

*dptra = (*sptra & (*sptra << 1 | (*sptra + 1) >> 31)

where *dptra represents a word pointer in the destination binary image and *sptra is the correspondent word pointer in the source binary image.

4.2. Stage 2: Fitness calculation

To calculate the overall fitness for all the pixels of the image two subsequent convolution procedures have to be accomplished. The compound SE is convolved with the cell-based SE which in its turn is convolved with the entire image. The optimized algorithm makes the first convolution explicit by calculating how many times the same pixel of the source image is considered within the convolution process. Or better, how many times an AND between same couple of pixels is computed. Therefore we just need to weight each SAND's element by a proper coefficient. Let \( n \geq 5 \) be the odd number representing the size of the cell-based SE. Then Eqs. 3 and 4 are the column vectors coefficients of the horizontal and of the vertical ANDs, respectively.

\[
\bar{u} = [1, 2, ..., 2, 1, 0]
\]

(3)

\[
\bar{w} = [1, 2, 3, ..., 3, 2, 1]
\]

(4)

For a given pixel \((x, y)\) of the original image, we define \( \hat{H} \) and \( \hat{V} \) as the component matrices of the SAND referring to the cell-based SE whose origin is translated onto the pixel \((x, y)\) (Eqs. 5 and 6).

\[
\hat{H} = [H_{y-\frac{n-1}{2}, i, \frac{x+1}{2}, \frac{x+1}{2}+1}], \quad i, j \in [0..n - 1]
\]

(5)

\[
\hat{V} = [V_{y-\frac{n-1}{2}, i, \frac{x+1}{2}, \frac{x+1}{2}+1}], \quad i, j \in [0..n - 1]
\]

(6)

In conclusion, the fitness value of the pixel \((x, y)\) for a given cell-based SE of size \( n \) inside the original image is expressed by Eq. 7:

\[
f_{x,y} = \bar{u}' \times \hat{H} \times \bar{u} + \bar{u}' \times \hat{V} \times \bar{w}
\]

(7)

As for the second convolution, most of the indexes for two consecutive pixel overlap for a certain extent. Also these overlapping configurations must be taken into consideration to speed up the overall process.

Regarding with the computational complexity, the original algorithm computational complexity is given by Eq. 8 while the optimized version computational complexity is given by Eq. 9, where \( w, h \) are the width and the height of the image respectively and \( N = w \times h \). Eq. 9 comes from the convolution in the vertical direction; by swapping \( h \) and \( w \) we achieve the computational complexity yielded by the convolution in the horizontal direction \((O(h \cdot n))\). Generally speaking, we can state that the computational complexity of the optimized algorithm is upper bounded by \( O(\sqrt{N} \cdot n) \) (in case \( h = w \)).

\[
N \cdot 12(n - 2)^2 \sim O(N \cdot n^2)
\]

(8)

\[
\left(n + \left(h - \frac{n - 1}{2}\right)\right) \sim O(w \cdot n)
\]

(9)

5. Experimental Results

The experiments accomplished on an AMD Athlon XP 2000 processor aim to evaluate the improvement of the optimized method SF in terms of scalability as the size \( n \) of the cell-based SE changes. Timing values have been averaged on all the gray level frames (384 × 288 in size) of
Figure 5: Time performance of the original and the optimized version of the SF method. The optimized algorithm scales linearly with the size of the cell-based SE.

a few hours long outdoor sequence. The average has been calculated just for reliability purposes, even though timing values practically do not change across frames and they do not depend even from the sequence. Fig. 5 shows how the optimized algorithm time performance scales linearly with \( n \) while the original one’s performance increases quadratically with \( n \). As for the absolute timing performance values of the optimized algorithm they range from \( 12 \text{ms} (n = 3) \) to \( 36 \text{ms} (n = 13) \). Experimental results accomplished on different sequences show that the optimized algorithm works at more than \( 35 \text{fps} \) (frame per second) for \( n = 9 \) (a commonly used value) while the original algorithm worked at less than \( 2 \text{fps} \). This means that while the old implementation could not be considered for real time sequence analyses, the optimized version is suitable to be employed in real time applications (actually, it is successfully embedded in our visual surveillance system).

6. Conclusion

In this work we present an optimized version of the SF method developed in [1]. Although the original algorithm is very effective in segmenting objects in very noises images it cannot be utilized for real time sequence analyses. The optimized version we developed reduces the overall computational complexity from \( O(N \cdot n^2) \) to \( O(\sqrt{N} \cdot n) \) (where \( N \) is the image size and \( n \) is the size of the cell-based SE of the SF method) thus making the SF method suitable for real time applications.

References


Combining Correlation Kernels and Kansei Information in Kernel CCA in Texture Classification

Yo Horikawa and Yujiro Ohnishi
Faculty of Engineering, Kagawa University
2217 Hayashi-cho, Takamatsu 761-0396, Japan
Email: horikawa@eng.kagawa-u.ac.jp

Abstract—The authors study effectiveness of combining correlation kernels in kernel canonical correlation analysis (kCCA) for the classification of texture images. Kansei information collected through questionnaires to the public with impression words for the images is also adopted. It is shown that classification performance increases with multiple correlation kernels and Kansei information in kCCA.

1. Introduction

Kernel methods have attracted much attention in pattern recognition [1], [2]. In the kernel methods, the inner product of feature vectors is replaced to a nonlinear kernel function. A nonlinear mapping of the feature vector to a high-dimensional space is then performed in an implicit manner. Kernel canonical correlation analysis (kCCA) is a kernel version of canonical correlation analysis [3]-[5]. Canonical correlation analysis was proposed by H. Hotelling in 1935 and finds linear transformations that yield maximum correlation between two kinds of features of objects.

The correlation kernel is an inner product of the autocorrelation functions of the feature vectors [6], [7]. The characteristic of the correlation kernel is that higher-order kernels are effectively calculated [8]. The correlation kernels are considered to be suitable to image data, which have strong spatial correlations. SVM and kCCA with the correlation kernels were applied to invariant texture classification [9], [10].

In this study we deal with combining correlation features and use of Kansei information in kCCA. First, combining classifiers and the fusion of information in pattern classification have been of wide interest [11], [12]. Combining classifiers can have higher classification performance than single classifiers. We apply it in kCCA by combining the canonical variates obtained with the correlation kernels of multiple-orders. Second, ‘Kansei’, the perceptual and cognitive ability to feel objects, has been of wide use in various fields of research. Kansei information was applied to image retrieval systems with CCA [13]. We adopt it in classification with kCCA as the second feature instead of a class indicator.

In the following, kCCA and correlation kernels are briefly explained in Sect. 2. In Sect 3, the method of obtaining and using Kansei information on images are described. Method and result of texture classification experiment are shown in Sect. 4. Discussion and conclusion are given in Sect. 5.

2. Kernel CCA and Correlation Kernels

2.1. Kernel Canonical Correlation Analysis

We mention a kernel version of canonical correlation analysis (kCCA) [3]-[5]. Let \((x_i, y_i), (1 \leq i \leq n)\) be a pair of the feature vectors of \(n\) sample objects, which describe different aspects of the objects, e.g., sounds and images. Define kernel matrices \(\Phi\) and \(\Theta\) by \(\Phi_{ij} = \phi(x_i, x_j)\) and \(\Theta_{ij} = \theta(y_i, y_j), (1 \leq i, j \leq n)\), which correspond to the inner products of implicit functions of \(x\) and \(y\), respectively. Then we obtain the eigenvectors \((f^+, g^+)\) of the generalized eigenproblem:

\[
\begin{bmatrix}
0 & \Phi & 0 \\
\Phi & 0 & \Theta \\
0 & \Theta & \gamma I
\end{bmatrix}
\begin{bmatrix}
f \\
g \\
g
\end{bmatrix}
= \lambda
\begin{bmatrix}
f \\
g \\
\gamma I
\end{bmatrix}
\]

Small multiples of the identity matrix \(\gamma I\) are added for the regularization. The canonical variates \(u\) and \(v\) of \((x, y)\) of a object are linear projections of the implicit functions of \(x\) and \(y\) which maximize correlation between them. They are obtained with the eigen vectors \(f\) and \(g\) by

\[
u = \sum_{i=1}^{n} \gamma I f_i \phi(x_i, x) \\v = \sum_{i=1}^{n} \gamma I g_i \theta(y_i, y)
\]

An indicator vector is used as the second feature vector \(y\) for classification problems [3]. When an object \(x\) is categorized into one of \(C\) classes, the indicator vector corresponding to \(x\) is defined by

\[
y = (y_1, \ldots, y_C)^T \quad \begin{cases} y_c = 1 & \text{if } x \text{ belongs to class } c \\ y_c = 0 & \text{otherwise} \end{cases} \quad (1 \leq c \leq C)
\]

Then the linear inner product \(y_i^T y_j\) is used as the kernel function \(\theta(y_i, y_j)\). The canonical variates \(u_i, (1 \leq i \leq C-1)\) are obtained corresponding to non-zero eigenvalues of Eq. (1). Standard classification methods, e.g., the nearest neighbor method, neural networks and SVM are applied in the canonical variate space.
2.2. Correlation Kernels

The autocorrelation of the original feature vector \( x \) is used in the correlation kernels [6], [7]. In the following, we consider 2-dimensional image data \( x(l, m) \), \( 1 \leq l \leq L \), \( 1 \leq m \leq M \) as the feature vector \( x \). The \( k \)-th order autocorrelation \( r_{x}(l_{1}, l_{2}, \ldots, l_{k}, m_{1}, m_{2}, \ldots, m_{k}) \) of \( x(l, m) \) is defined by

\[
r_{x}(l_{1}, l_{2}, \ldots, l_{k}, m_{1}, m_{2}, \ldots, m_{k}) = \sum_{l} \sum_{m} x(l, m)x(l+l_{1}, m+m_{1}) \ldots x(l+l_{k}, m+m_{k})
\]

(4)

The inner product of the autocorrelations \( r_{x} \) and \( r_{y} \) of image data \( x(l, m) \) and \( y(l, m) \) is calculated by the sum of the \( k \)-th power of the cross-correlation \( c_{x_{i}y_{j}}(l_{1}, m_{1}) \) of the image data

\[
r_{x_{i}y_{j}} = \sum_{l=0}^{L} \sum_{m=0}^{M} (c_{x_{i}y_{j}}(l_{1}, m_{1}))^{k}/(L,M)
\]

(5)

\[
c_{x_{i}y_{j}}(l_{1}, m_{1}) = \sum_{l=1}^{L} \sum_{m=1}^{M} x(l, m)x(l+l_{1}, m+m_{1})/(LM)
\]

(6)

where \( L \) and \( M \) are the ranges of the lags \( l \) and \( m \) of the cross-correlation [8]. Computational costs are reduced in the practical order even for high-order \( k \) of the correlation and large data size \( L \) and \( M \) since the calculation of the explicit values of the autocorrelations are avoided. Equation (5) is employed as the \( k \)-th order correlation kernel function \( k(x_{i}, x_{j}) \) and thus \( \Phi \) in Eq. (1).

Since the performance of the correlation kernels of odd or higher-orders is degraded, the following modified versions of the correlation kernels have been proposed [14] and will be used in this study.

\[
L_{p} \text{ norm kernel (P)}
\]

\[
r_{x_{i}y_{j}} = \text{sgn}(c_{x_{i}y_{j}}(l_{1}, m_{1})) \sum_{i=m_{1}}^{M-1} |c_{x_{i}y_{j}}(l_{1}, m_{1})|^{k}/(M)
\]

(7)

Absolute correlation kernel (A)

\[
r_{x_{i}y_{j}} = \sum_{i=m_{1}}^{M-1} |c_{x_{i}y_{j}}(l_{1}, m_{1})|^{k} \]

(8)

Max norm kernel (Max)

\[
r_{x_{i}y_{j}} = \max_{i=m_{1}}^{M-1} c_{x_{i}y_{j}}(l_{1}, m_{1})
\]

(9)

The \( L_{p} \) norm and absolute correlation kernels take the \( k \)th roots and absolute values, respectively, of the original ones. The max norm kernel is regarded as the \( L_{p} \) norm kernel in the limit of \( k \to \infty \).

3. Collection and Use of Kansei Information

We made a questionnaire system on Web to collect Kansei information from the public. Java Server Pages (JSP) was used for the client-server system. Figure 1 shows a Web page for the questionnaire, in which people are asked to answer questions in Japanese. The page shows 30 texture images arbitrarily taken from the Brodatz album [15] and asks people to choose one image which matches each Japanese impression word. Twenty impression words for images [16] are used: delicate, beautiful, bold, sharp, decorative, fine, sophisticated, simple, soft, deep, impressive, quiet, elegant, chic, natural, hard, grave, silent, solid, rural. We sent requests for the questionnaire by e-mail to undergraduate students in our faculty and obtained answers from 50 students.

As a Kansei feature of each image, we use a 20-dimensional vector the element of which is the number of the vote of the impression word to the image, which we call an impression vector. For instance, the vector for the texture image No. 1 (Brodatz D101), top-left in Fig. 1, is

\[
y_{1} = (1,1,0,7,1,25,1,0,7,2,1,0,1,0,0,3,0,2)
\]

(10)

which indicates that many answerers thought the image simple. We will use the impression vector as the second feature vector \( y \) in kCCA in Sect. 4.

4. Experiment in Texture Classification

4.1. Method

Experiment on the classification of texture images with kCCA using the correlation kernels and Kansei information was done. Thirty texture images in the Brodatz album shown in Fig. 1 are used. The 8bit image data of 320×320 pixels are obtained from AMOVIP-DB [17]. Twenty subimages of 50×50 pixels are taken from each original image without overlap and one half (ten) of them, 300 images in total, is used as sample data and the other half is used as test data.

As the second feature vector \( y \) in kCCA , we use the indicator vectors (Eq. (3)) or the impression vectors explained in Sect. 3, in which Kansei information is taken into account. The eigenvectors \( f_{i} \) (\( 1 \leq i \leq n_{c} \)) corresponding to the non-zero eigenvalues of Eq. (1) are calculated with the sample data. The number \( n_{c} \) of the elements is 29 for the indicator vectors and 19 for the impression vectors. The corresponding canonical variates \( u_{i} \) (\( 1 \leq i \leq n_{c} \)) of the images are calculated by Eq. (2). To avoid numerical instability \( \Phi_{y} \) are divided by \( E(\Phi_{y}) \) in Eq. (1) and the values of \( \gamma_{x} \) and \( \gamma_{y} \) are set to be 0.1n. The lags
$L_1$ and $M_1$ of the correlations in Eq. (5) are taken to be 10. To express the set $(\phi, y)$ of the kernel function $\phi$ and the second feature vector $y$, the following symbols are used.

the $k$th-order correlation kernel: $C_k$, the $k$th-order $L_p$ norm kernel: $P_k$, the $k$th-order absolute correlation kernel: $A_k$, the max norm kernel: Max, for the kernel function $\phi$ of image data;

the indicator vector: $I$, the impression vector $K$, for the second feature vector $y$.

A simple nearest neighbor method (1-NN) in the canonical variate space is used for the classification of the test data. As combined features, Cartesian spaces of the canonical variates of kCCAs with various sets of the kernel functions and the second feature vectors are used, e.g., $(u_1, \ldots, u_{nc}, u_1', \ldots, u_{nc}')$. Classification is also done with 1-NN in the Cartesian space. It can be shown that appropriate choice of the canonical variates gives higher correct classification rates (CCRs) than the use of the all canonical variates in most cases. Then CCRs are calculated with one to $nc$ canonical variates taken in the order of the values of the corresponding eigenvalues, namely from $(u_1)$ to $(u_1, \ldots, u_{nc})$ in each case and the highest CCR is shown in the results.

Further, a voting method is also tested for combining more than three kCCAs. It is because the number of the combinations of the canonical variates of kCCAs in obtaining the optimal set of them increases intractably. As mentioned above CCRs tend to decrease when the all canonical variates are combined.

### 4.2. Result

#### 4.2.1. Canonical Variates

The scatter diagrams of the canonical variates $(u_1, u_2)$ of the first and second largest eigenvalues in kCCA with the 2nd-order correlation kernels $(C_2, I)$ and $(C_2, K)$ are shown in Fig. 2. The centroids of 10 sample images of the same texture images in the canonical variates of the six largest eigenvalues are shown. Figure 3 shows the dendrograms of the texture images obtained with agglomerative hierarchical clustering with average linkage. Use of Kansei information $(C_2, K)$ causes the scatter diagram and dendrogram different from those with the indicator vectors $(C_2, I)$. It is then expected that the canonical variates with Kansei information work as complementary features in classification tasks.

#### 4.2.2 Classification Performance

Table 1 shows the correct classification rates (CCRs) with kCCAs. The use of the indicator vectors $I$ gives higher CCRs than the expression vectors $K$ for the all kernel functions $(C_2, \text{Max}, A_1, P_4)$, which confirms that the indicator vectors are suitable for classification tasks.

Table 2 shows CCRs with the combination of the canonical variates of two different kCCAs. The highest CCR is shown with boldface in each raw. The combination of the different second feature vectors $(I$ and $K)$ can show the highest CCR, e.g., $(C_2, K)$ and $(\text{Max}, K)$. It is also shown that the combination of the best single kCCAs $(C_2, I)$ and $(A_1, I)$ does not give the highest CCR.
Table 3 shows CCRs with the combination of three kCCAs, one of which is (C2, I). Higher CCRs are still obtained than two kCCAs. It can be shown that CCR increases to 0.627 with the combination of four kCCAs, e.g. (C2, I or K), (Max, I), (A1, I or K), (P4, I or K).

Further, CCRs obtained with the voting of the 1-NN classification results with kCCAs are shown in Table 4. The results of kCCAs are added in the voting from left to right. The CCR increases from 0.517 of (C2, I) to 0.597 of the voting of nine kCCAs ((C2, I), ···. (A3, I)). However, the voting of more kCCAs degrades CCRs slightly.

Adding the correlation variates of kCCAs up to four always increased CCRs in the experiment. However, increases in the number of kCCAs may decrease CCRs as was shown in Table 4. This is ascribed to correlations between the canonical variates of various correlation kernels.

Table 1 CCR with kCCA.

<table>
<thead>
<tr>
<th>(C2, I)</th>
<th>(Max, I)</th>
<th>(A1, I)</th>
<th>(P4, I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.517</td>
<td>0.503</td>
<td>0.537</td>
<td>0.510</td>
</tr>
</tbody>
</table>

Table 2 CCR with two kCCAs.

<table>
<thead>
<tr>
<th>(C2, I)</th>
<th>(Max, I)</th>
<th>(A1, I)</th>
<th>(P4, I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.580</td>
<td>0.547</td>
<td>0.537</td>
<td>0.500</td>
</tr>
<tr>
<td>(Max, I)</td>
<td>0.507</td>
<td>0.550</td>
<td>0.563</td>
</tr>
<tr>
<td>(A1, I)</td>
<td>0.547</td>
<td>0.563</td>
<td>0.550</td>
</tr>
<tr>
<td>(P4, I)</td>
<td>0.547</td>
<td>0.550</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Table 3 CCR with three kCCAs (incl. (C2, I)).

<table>
<thead>
<tr>
<th>(C2, I)</th>
<th>(Max, I)</th>
<th>(A1, I)</th>
<th>(P4, I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.580</td>
<td>0.547</td>
<td>0.537</td>
<td>0.500</td>
</tr>
<tr>
<td>(Max, I)</td>
<td>0.507</td>
<td>0.550</td>
<td>0.563</td>
</tr>
<tr>
<td>(A1, I)</td>
<td>0.547</td>
<td>0.563</td>
<td>0.550</td>
</tr>
<tr>
<td>(P4, I)</td>
<td>0.547</td>
<td>0.550</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Table 4 CCR with voting of kCCAs with multiple correlation kernels.

<table>
<thead>
<tr>
<th>(C2, I)</th>
<th>(Max, I)</th>
<th>(A1, I)</th>
<th>(P4, I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.580</td>
<td>0.547</td>
<td>0.537</td>
<td>0.500</td>
</tr>
<tr>
<td>(Max, I)</td>
<td>0.507</td>
<td>0.550</td>
<td>0.563</td>
</tr>
<tr>
<td>(A1, I)</td>
<td>0.547</td>
<td>0.563</td>
<td>0.550</td>
</tr>
<tr>
<td>(P4, I)</td>
<td>0.547</td>
<td>0.550</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Table 3 with three kCCAs (incl. (C2, I)).

<table>
<thead>
<tr>
<th>(C2, I)</th>
<th>(Max, I)</th>
<th>(A1, I)</th>
<th>(P4, I)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.580</td>
<td>0.547</td>
<td>0.537</td>
<td>0.500</td>
</tr>
<tr>
<td>(Max, I)</td>
<td>0.507</td>
<td>0.550</td>
<td>0.563</td>
</tr>
<tr>
<td>(A1, I)</td>
<td>0.547</td>
<td>0.563</td>
<td>0.550</td>
</tr>
<tr>
<td>(P4, I)</td>
<td>0.547</td>
<td>0.550</td>
<td>0.563</td>
</tr>
</tbody>
</table>

5. Discussion

In this study, multiple correlation kernels were combined and Kansei information was employed in kCCA. The experiment of texture classification showed that the classification performance increases by combining the canonical variates of different correlation kernels. Further, Kansei information gives different features and contributes to increases in the performance of kCCA.

Optimal choice of the canonical variates is necessary to obtain the best performance. The optimal numbers of the canonical variates of kCCAs differ from each other. Further, combining the optimal canonical variates of single kCCAs does not usually show the highest CCR in the combination of them. The choice was done empirically in the experiment and it is a rather formidable task when the number of kCCAs increases. A simple voting of kCCAs also makes CCRs high, but lower than the optimal combination of the canonical variates.

References

Comparing Two-Layer CNN with van der Pol Oscillators Coupled by Inductors

Koji Urabe† and Yoshifumi Nishio‡

†Department of Electronic Control Engineering, Niihama National College of Tech.
7-1 Yagumo, Niihama, Ehime 792-8580, JAPAN
‡Dept. of Electrical and Electronic Engineering, Tokushima University
2-1 Minami-Josanjima, Tokushima 770-8506, JAPAN
Email: koji@ect.niihama-nct.ac.jp, nishio@ee.tokushima-u.ac.jp

Abstract—In this work, we present that the 2-layer CNN is similar to van der Pol oscillators coupled by inductors, which can generate the phase-wave propagation phenomena. However, we cannot observe the phase-wave propagation phenomena in the original 2-layer CNN. Therefore, we introduce a modified 2-layer CNN. We clearly show the correspondence between the modified 2-layer CNN and the van der Pol oscillators coupled by inductors.

1. Introduction

Cellular Neural Networks (CNN) were invented by L.O.Chua and L.Yang in 1988 [1]. CNN are constructed by cells connected each other. The cell contains linear and nonlinear current sources controlled by voltage. Already a lot of applications and VLSI implementations of CNN were reported. Many nonlinear phenomena such as pattern formation and autowaves could be observed in CNN. Investigating the nonlinear phenomena is an important work for clarifying dynamics of CNNs. On the other hand, phase-wave propagation phenomena in van der Pol oscillators coupled by inductors were reported [2][3]. It is known that 2-layer CNNs can exhibit phase-wave propagation phenomena by choosing an appropriate set of the parameters [4][5].

In this work, we report the detailed investigation on the similarity between the 2-layer CNN and the van der Pol oscillators coupled by inductors. In particular we make clear the correspondence of the parameters of the two systems. Further, we confirm that slight difference between them prevents to generate some of phase-wave propagation phenomena.

2. van der Pol Oscillators Coupled by Inductors

The van der Pol oscillators coupled by inductors $L_0$ as a comparative object used in this study are shown in Fig. 1.

The circuit equations governing the circuit in Fig. 1 are written as:

\[
\frac{du_k}{dt} = w_k \quad \quad (2)
\]

\[
\frac{dw_k}{dt} = -u_k + \alpha (u_{k+1} - 2u_k + u_{k-1}) + \epsilon (w_k - w_k^2/3) \quad \quad (3)
\]

where

\[
t = \sqrt{L_1 C}, \quad u_{i,k} = \sqrt{\frac{C g_1}{3 L_1 g_3}} u_k, \quad v_k = \sqrt{\frac{g_1}{3 g_3}} w_k.
\]

\[
\alpha = \frac{L_1}{L_0}, \quad \epsilon = g_1 \sqrt{\frac{L_1}{C}}. \quad \quad (4)
\]

And, we consider the boundary conditions as follows:

\[
u_0 = u_1, \quad u_{N+1} = u_N. \quad \quad (5)
\]

It should be noted that $\alpha$ corresponds to the coupling of the oscillators and $\epsilon$ corresponds to the nonlinearity of the oscillators. In this study, we calculate (2) and (3) by using the Runge-Kutta-Gill method.

In this study, we use $N = 8$, $\alpha = 0.050$ and $\epsilon = 0.30$ for numerical analysis. And the initial conditions are given as follows:

1. Setting the initial conditions of all oscillators as the same.
2. Putting the arbitrary phase difference of the voltage and the current to one oscillator.

Putting the phase difference $+180$ [deg] to the 1st oscillator, simulation result is shown in Fig. 2. In Fig. 2, the vertical axis is the sum of voltages of adjacent oscillators, and horizontal axis is time. If the sum of voltages of adjacent oscillators is zero, the phase difference between adjacent oscillators is $+180$ [deg]. At first, only the 1st oscillator has phase difference $+180$[deg], phase difference propagates to the adjacent oscillator as time goes.
3. Two-Layer CNN

In this study, we use 1-dimensional 2-layer CNN as shown in Fig. 3.

![Figure 3: One-dimensional 2-layer CNN.](image)

The circuit equations governing the CNN in Fig. 3 are written as

\[
\begin{align*}
    x_{1,k} &= -x_{1,k} + a_1 x_{1,k} + c_1 x_{2,k} + d_1 y_{2,(k-1)} + d_2 y_{2,(k+1)} \quad (6) \\
    x_{2,k} &= -x_{2,k} + a_2 y_{2,k} + c_2 y_{1,k} + d_2 y_{1,(k-1)} + d_3 y_{1,(k+1)} \quad (7) \\
    y_{f,k} &= f(x_{f,k}) = 0.5([x_{1,k} + |x_{1,k}|] - [x_{2,k} - |x_{2,k}|]) \quad (8)
\end{align*}
\]

where \( x_{1,k} \) is the state, \( y_{f,k} \) is the output of \( \text{CELL}_{f,k} \), \( a_f, c_f, \) and \( d_f \) are the feedback parameters from the output of its own cell, from the output of the cell which is at the same position in the other layer, and from the output of the neighborhood cell in the other layer, respectively.

In order to investigate the correspondence between the 2-layer CNN and the van der Pol oscillators, we replace the output feedbacks \( y_{f,k} \) in Eqs. (6) and (7) by the state feedbacks \( x_{f,k} \), except \( y_{2,k} \), as follows:

\[
\begin{align*}
    x_{1,k} &= -x_{1,k} + a_1 x_{1,k} + c_1 x_{2,k} + d_1 x_{2,(k-1)} + d_1 x_{2,(k+1)} \quad (9) \\
    x_{2,k} &= -x_{2,k} + a_2 x_{2,k} + c_2 x_{1,k} + d_2 x_{1,(k-1)} + d_2 x_{1,(k+1)} \quad (10)
\end{align*}
\]

By comparing Eq. (9) with Eq. (2), some of the template values are decided as follows:

\[
\begin{align*}
    a_1 &= -1, \quad c_1 = 1, \quad d_1 = 0. \quad (11)
\end{align*}
\]

Further, we approximate the piecewise linear function \( f(\cdot) \) of the output feedback from its own cell by the 3rd order polynomial expression.

\[
\begin{align*}
    -x_{2,k} + a_2 y_{2,k} &= -x_{2,k} + 0.5 a_2 ([x_{2,k} + 1] - [x_{2,k} - 1]) \quad (12) \\
    &\sim \epsilon \left( x_{2,k} - x_{2,k}^3 / 3 \right) \quad (13)
\end{align*}
\]

Equation (13) with \( \epsilon = 0.30 \) is shown by a dotted line in Fig. 4. We approximate this curve by Eq. (12) with \( a_2 = 1.2 \), whose curve is shown by a solid line in Fig. 4.

![Figure 4: Approximation of piecewise linear function.](image)

Equation (10) can be rewritten using the Eqs. (12) and (13) as follows:

\[
\begin{align*}
    x_{2,k} &= \epsilon \left( x_{2,k} - x_{2,k}^3 / 3 \right) + c_2 x_{1,k} + d_2 x_{1,(k-1)} + d_2 x_{1,(k+1)} \quad (14)
\end{align*}
\]

By comparing Eq. (14) with Eq. (3), the other template values are decided as follows:

\[
\begin{align*}
    a_2 &= 1.2, \quad c_2 = -(1 + 2 \alpha), \quad d_2 = \alpha \quad (15)
\end{align*}
\]

Using the template values in Eqs. (11) and (15), Eqs. (9) and (10) are rewritten as follows:

\[
\begin{align*}
    x_{1,k} &= x_{2,k} \quad (16) \\
    x_{2,k} &= -x_{1,k} + \alpha (x_{1,(k-1)} - 2 x_{1,k} + x_{1,(k+1)}) + \epsilon \left( x_{2,1} - x_{2,1}^3 / 3 \right) \quad (17)
\end{align*}
\]
Now, we consider the boundary conditions as follows:

\[ x_{1,0} = x_{1,1} \quad (18) \]
\[ x_{1,N+1} = x_{1,N} \quad (19) \]

Equations (16)-(19) are completely the same as the equations of the van der Pol Oscillators coupled by the inductors (2), (3), and (5).

We expect that the phase-wave propagation phenomena, which are observed in coupled van der Pol oscillators, can be generated in the 2-layer CNN (6) and (7) with the template values (11) and (15). Namely,

\[ x_{1,k} = y_{2,k} \quad (20) \]
\[ x_{2,k} = -y_{1,k} + \alpha(y_{1,(k-1)} - 2y_{1,k} + y_{1,(k+1)}) + \epsilon(x_{2,1} - x_{2,1}^3/3) \quad (21) \]

However, the simulated results of Eqs. (20) and (21) with \( \alpha = 0.050 \) and \( \epsilon = 0.30 \) did not show the continuously existing phase-wave propagation phenomena. Figure 5 shows an example of the observed phenomena for \( N = 8 \) obtained by using the Runge-Kutta-Gill Method.

![Figure 5: Simulated result of 2-layer CNN (N = 8).](image)

The difference between Eqs. (20) and (21) and Eqs. (2) and (3) is only the feedbacks. Namely, the output feedbacks in the 2-layer CNN and the state feedbacks in the coupled van der Pol oscillators.

### 4. Modified 2-Layer CNN

We introduce the modified 2-layer CNN that the output feedbacks from the cell at the same position in the other layer are replaced by the state feedback. The equations governing the modified 2-layer CNN are written as follows;

\[ x_{1,k} = -x_{1,k} + a_1 y_{1,k} + c_1 x_{2,k} \quad (22) \]
\[ x_{2,k} = -x_{2,k} + d_2 y_{2,k} + c_2 x_{1,k} + d_2 y_{1,(k-1)} + d_2 y_{1,(k+1)} \quad (23) \]

Equations (22) and (23) are rewritten using the templates (11) and (15);

\[ x_{1,k} = x_{2,k} \quad (24) \]
\[ x_{2,k} = -x_{1,k} + \alpha(y_{1,(k-1)} - 2y_{1,k} + y_{1,(k+1)}) + \epsilon(x_{2,1} - x_{2,1}^3/3) \quad (25) \]

The simulation result of Eqs. (24) and (25) is shown in Fig. 6. In Fig. 6, we can observe the phase-wave propagation phenomena similar to Fig. 2.

![Figure 6: Simulated result of modified 2-layer CNN (N = 8).](image)

From this result we consider that the cells at the layer 1 correspond to the differentiation of \( u_k \). The output feedbacks from their own cells at the layer 2 correspond to the nonlinear resistors, and the output feedbacks from the cells at the layer 1 correspond to the couplings between the oscillators.

### 5. Phase-Wave Propagation Phenomena

The simulated result of the coupled van der Pol oscillators with different coupling parameter \( \alpha = 0.10 \) is shown in Fig. 7. The speed of the phase-wave propagation is faster than Fig. 2 with \( \alpha = 0.05 \). Namely, the propagation speed can be controlled by changing the coupling inductor \( L_0 \) in the coupled oscillators.

We can confirm similar effect in the modified 2-layer CNN by changing the template values. From \( \alpha = 0.10 \),

\[ c_2 = -(1 + 2\alpha) = 1.2, \quad d_2 = \alpha = 0.10. \]

The simulated result of the modified 2-layer CNN with the calculated template values is shown in Fig. 8. We can see that the speed of the phase-wave propagation is faster than Fig. 6 similar to Fig. 7.

Figure 9 shows some typical examples of the phase-wave propagation phenomena in the van der Pol oscillators coupled by the inductors.

The corresponding simulated results by the modified 2-layer CNN are shown in Fig. 10. We can observe the reflection of the waves, but cannot observe the extinction by the collision.

![Figure 7: Simulated result of van der Pol oscillators with different \( \alpha = 0.10 \).](image)
In this study, we clearly explained the correspondence between an array of van der Pol oscillators coupled by inductors and 1-dimensional 2-layer CNN. By investigating the generated phase-wave propagation phenomena, we found that the difference between the output feedbacks and the state feedbacks played an important role to make difference in the observed phenomena.

In future works, we will try to clarify the mechanism of the phase-wave propagation phenomena by investigating the modified 2-layer CNN, so that we can control the phase-wave propagation for applications.

References


Figure 8: Simulated result of modified 2-layer CNN with different template values. $a_1 = 1$, $c_1 = 1$, $d_1 = 0$, $a_2 = 1.2$, $c_2 = -1.2$, and $d_2 = 0.10$.

Figure 9: Examples of phase-wave propagation phenomena in van der Pol oscillators coupled by the inductors. $N = 15$, $a = 0.10$ and $\epsilon = 0.30$.

Figure 10: Example of phase-wave propagation phenomena in modified 2-layer CNN, $N = 15$, $a_1 = 1$, $c_1 = 1$, $d_1 = 0$, $a_2 = 1.2$, $c_2 = -1.2$, and $d_2 = 0.10$. 

6. Conclusions

(a) Extinction by collision of two waves.

(b) Reflection by collision of two waves.

(a) Extinction by collision can not be observed.

(b) Reflection by collision of two waves.
Study on the number of equilibrium points
for a 1-D DBCNN with no input

Hajime HARA†, Tetsuo NISHI†† and Norikazu TAKAHASHI†††

†Department of Electronics and Photonic Systems Engineering, Hiroshima Institute of Technology
2-1-1 Miyake, Saeki-ku, Hiroshima 731-5193 Japan
††Faculty of Science and Engineering, Waseda University
3-14-9 Okubo, Shinjuku-ku, Tokyo 169-0072 Japan
†††Department of Computer Science and Communication Engineering, Kyushu University
6-10-1 hakoizaki, higashi-ku, Fukuoka, 812-8581 Japan
Email: hara@cc.it-hiroshima.ac.jp, nishi-t@waseda.jp

Abstract—We investigate the number of equilibrium points for a 1-D DBCNN (one-dimensional discrete-time binary cellular neural network) with no input and give explicit formulae for them. We see the number of equilibrium points is fairly small.

1. Introduction

The most fundamental problems of cellular neural networks(CNN) from the theoretical points of view are the stability and the number of equilibrium points, while the ability for 1- and 2-dimensional signal processings is also quite important from the practical point of view [1]-[4]. There still remain unsolved problems even for one-dimensional CNN’s [5]-[7].

Concerning the stability conditions of a 1-dimensional discrete-time binary cellular neural network (abbreviated as a 1-D DBCNN), Sato et al.[7] gave the necessary and sufficient conditions in terms of changeable sets and Hara et al.[9] and Nishi et al.[8] recently gave the necessary and sufficient conditions in terms of system parameters for both no input case and nonzero input case. On the other hand Thiran et al[6] studied on 1-dimensional analog CNNs and gave the stability conditions for them. So the stability problem is considered to be solved for a kind of 1-dimensional systems with one-neighborhood connection.

Thiran et al[6] also gave explicitly the number of equilibrium points in saturated regions of the above systems. The number of equilibrium points is important not only theoretically but also practically since it is closely related to the variety of signal processings (e.g., as memory).

In this paper we study on the number of equilibrium points for a 1-D DBCNN with no input and give explicit formulae for it. The discussion is based on the previous results described by changeable sets (which will be explained later). From these formulae we see that the number of equilibrium points is fairly small.

2. Preliminaries

The behavior of a 1-D DBCNN denoted by $S$ can generally be described by the equation:

$$x(k+1) = \text{sgn}[Ax(k) + Bu + \theta]$$

where $x(k) = [x_1(k), \cdots, x_n(k)]^T$ and $u = [u_1, u_2, \cdots, u_n]^T$ are a binary state vector at time $k$ and a binary time-invariant input vector respectively, $n$ is the dimension of $S$, $A$ and $B$ are $n \times n$ matrices determined by the A- and B-templates, $\theta$ is a scalar representing the threshold value, and 1 is an n-dimensional column vector consisting of 1 only. In particular $x(0)$ is the initial state vector, which can be used as another input data in many applications. We assume the 1-neighborhood DBCNN. Then Eq.(1) can be rewritten in a scalar form as:

$$x_i(k+1) = \text{sgn}[\beta x_{i-1} + \alpha x_i + \gamma x_{i+1}(k) + \theta]$$

where $x_i(k)$ and $x_{i+1}(k)$ are constants independent of $k$.

**Definition 1:** A 1-D DBCNN $S$ is said to be stable, if no limit cycle occur for any $x(0)$, any $u$, any boundary conditions on $x$ and $u$, and any value of the dimension $n$. The 1-D DBCNN being not stable are said to be unstable.

Since we consider the case of $u = 0$, the system equation can be written as follows:

$$x_i(k+1) = \text{sgn}[\beta x_{i-1} + \alpha x_i + \gamma x_{i+1}(k) + \theta]$$

Then the problem is stated as follows: Prescribed parameters $\alpha$, $\beta$, $\gamma$ and $\theta$, how many equilibrium points does the system $S$ have? We will give the answer to the above problem.  

In the case of cellular automata the “sgn” function in Eq.(1) should be replaced with an arbitrary logic function of $x(k)$. 

727
3. Previous results and definitions

3.1. Stability conditions in terms of changeable sets

The 3-tuple \((x_{i-1}(k), x_i(k), x_{i+1}(k))\) takes one of the following eight patterns: 
\((-,-,-), (-,-, +), (-,+, -), (+,-,-), (+,-, +), (+,+, -),\) and 
\((+,+, +),\) where “+” and “−” mean +1 and −1, respectively. For some of these eight 3-tuples, 
\(x_i(k+1)\) changes from \(x_i(k)\) and for other 3-tuples \(x_i(k+1)\) is the same as \(x_i(k)\).

**Definition 2:** We call the former 3-tuples “changeable patterns” and the latter ones “invariant patterns”.

If \(\phi_1, \phi_2, \cdots, \phi_s\) are all of changeable patterns of \(S\), then we call \(\Phi \equiv \{\phi_1, \phi_2, \cdots, \phi_s\}\) a changeable set. Note that \(\Phi\) is independent of \(n\) and \(x(0)\). Since a changeable set uniquely determines the behavior of \(S\), we identify \(\Phi\) with \(S\). For example, we can say that “\(\Phi\) is stable” instead of “\(S\) is stable”.

Throughout this paper the variables \(y_i\) and \(y'_i(i = 1, 2, \cdots)\) denote the binary values 1 or −1 and “\(\overline{y}_i\)” means

\[
\overline{y}_i = \begin{cases} 
1 & \text{if } y_i = -1 \\
-1 & \text{if } y_i = 1 
\end{cases}
\] (4)

Then the previously obtained fundamental results are stated as follows:

**Theorem 1:** The system \(S\) described by Eq.(3) is unstable if and only if at least one of Eqs.(5) and (6) holds for some \(y_i\).

\[
\Phi \supseteq \{(y_1, y_2, y_3), (y_1, \overline{y}_2, y_3)\} \quad (5)
\]

\[
\Phi \supseteq \{(y_1, y_2, y_3), (y_1, \overline{y}_2, \overline{y}_3), (\overline{y}_1, y_2, \overline{y}_3)\} \quad (6)
\]

3.2. \(\alpha\)- and \(\bar{\alpha}\)-terms

We see from Eq.(3) that the behavior of the system can be determined only by the values of \(\pm|\beta| \pm \alpha \pm |\gamma| + \theta\). We classify these eight values into two as follows:

**Definition 3:** We call the terms \(\pm|\beta| + \alpha \pm |\gamma| + \theta\) \(\alpha\)-terms and \(\pm|\beta| - \alpha \pm |\gamma| + \theta\) \(\bar{\alpha}\)-terms respectively.

3.3. Regions \(A-i\) and \(\bar{A}-j\)

We will briefly state the relation between changeable sets and connection coefficients. This is easily done by drawing figures shown below.

When we arrange the \(\alpha\)-terms \(\pm|\beta| + \alpha \pm |\gamma|\) in the order of values (see Table 1), the largest (resp., smallest) is apparently \(|\beta| + \alpha + |\gamma|\) (resp., \(-|\beta| + \alpha - |\gamma|\)) and the second largest is \(|\beta| + \alpha - |\gamma|\) or \(-|\beta| + \alpha + |\gamma|\).

Without loss of generality we assume that

\[
|\beta| \geq |\gamma| \quad \text{i.e.,} \quad |\beta| + \alpha - |\gamma| \geq -|\beta| + \alpha + |\gamma| \quad (7)
\]

Similarly we arrange the \(\bar{\alpha}\)-terms \(\pm|\beta| - \alpha \pm |\gamma|\) in the order of values (see Table 2).

<table>
<thead>
<tr>
<th>(\theta)</th>
<th>(\alpha)-terms</th>
<th>Region (A-i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>\beta</td>
<td>+ \alpha +</td>
</tr>
<tr>
<td>(</td>
<td>\beta</td>
<td>+ \alpha -</td>
</tr>
<tr>
<td>(-</td>
<td>\beta</td>
<td>+ \alpha +</td>
</tr>
<tr>
<td>(-</td>
<td>\beta</td>
<td>+ \alpha -</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(\theta)</th>
<th>(\bar{\alpha})-terms</th>
<th>Region (\bar{A-j})</th>
</tr>
</thead>
<tbody>
<tr>
<td>(</td>
<td>\beta</td>
<td>- \alpha +</td>
</tr>
<tr>
<td>(</td>
<td>\beta</td>
<td>- \alpha -</td>
</tr>
<tr>
<td>(-</td>
<td>\beta</td>
<td>- \alpha +</td>
</tr>
<tr>
<td>(-</td>
<td>\beta</td>
<td>- \alpha -</td>
</tr>
</tbody>
</table>

Table 1 \(\alpha\)-terms

Table 2 \(\bar{\alpha}\)-terms

In Table 1 (resp., Table 2) \(\theta\)-axis can be divided into five regions denoted by \(A-i\) (resp., \(\bar{A}-j\)) \((i = 0, 1, \cdots, 4)\) by \(\pm|\beta| + \alpha \pm |\gamma|\) (resp., \(\pm|\beta| - \alpha \pm |\gamma|\)). The third columns in Tables are for later explanation.

3.4. S-point and T-point

**Definition 4:** If a changeable set \(\Phi\) satisfies

\[
\Phi \supseteq \{(y_1, y_2, y_3), (y_1, \overline{y}_2, \overline{y}_3)\} \quad (8)
\]

or

\[
\Phi \supseteq \{(y_1, y_2, y_3), (\overline{y}_1, y_2, \overline{y}_3)\} \quad (9)
\]

for some \(y_i\) \((i = 1, 2, 3)\), we say that \(S\) possesses an S-point and a T-point, respectively. Similarly, if \(\Phi\) satisfies Eq.(5) for some \(y_i\), then we say \(S\) has an S-T-point. We can easily verify that if an S-T-point exists, then both an S-point and a T-point inevitably occur. Then Theorem 1 can be rewritten as follows:

**Theorem 2:** The system \(S\) is unstable if and only if it has both an S-point and a T-point.

Existence of S- and T-points as well as \(\Phi\) can easily be seen from Tables 1 and 2. We will explain it by using examples.

3.5. Determination of \(\Phi\)

For example, we consider Region \(A-2\), which means from Table 1 that

\[
|\beta| + \alpha - |\gamma| > \theta > -|\beta| + \alpha + |\gamma| \quad (10)
\]

i.e.,

\[
|\beta| + \alpha - |\gamma| + \theta > 0, \quad -|\beta| + \alpha + |\gamma| + \theta < 0 \quad (11)
\]
This imply that for Region $A-2$
$$\Phi \supseteq \{(-, +, +), (-, +, -)\} \quad (12)$$
but $\Phi \not\supseteq \{(+, +, -)\}$.  
Similarly for Region $\bar{A}-3$, which means from Table 2 that
$$-|\beta| - \alpha + |\gamma| > -\theta > -|\beta| - \alpha - |\gamma| \quad (13)$$
i.e.,
$$-|\beta| - \alpha + |\gamma| + \theta > 0, \quad -|\beta| - \alpha - |\gamma| + \theta < 0 \quad (14)$$
This imply that for Region $A-3$
$$\Phi \supseteq \{(-, -, +), (+, -, -), (+, +, +)\} \quad (15)$$
but $\Phi \not\supseteq \{(-, -, -)\}.$

4. Investigation of 25 combinations

Previously we defined Regions $A-i$ ($i = 0, 1, \ldots, 4$) for $\alpha$-terms and Regions $A-j$ ($j = 0, 1, \ldots, 4$) for $\beta$-terms. For our purpose we investigate whether there exist both an $S$- and a $T$-point for each combination of $A-i$ and $A-j$ ($i, j = 0, 1, \ldots, 4$). Concretely, 1) is the sistem stable or not? 2) If stable, what is the number of equilibrium points?

We will exhaustively investigate all 25 combinations. This method is not effective but simple and easy to understand. To simplify the investigation, We can find that the Table 3 is symmetric on the diagonal. So we can investigate only 15 cases. The number of equilibrium point depends on the fixed boundary values $x_0(0)$ and $x_{n+1}(0)$. So we denote the number by $N(n; x_0(0), x_{n+1}(0))$. If $N$ depends on neither $x_0(0)$ nor $x_{n+1}(0)$, we simply write it $N(n)$.

Remark 1: If the region described by both $A-i_0$ and $A-j_0$ is unstable, then the regions described by both $A-i$ and $A-j$ $i \geq i_0$ and $j \geq j_0$ are unstable.

<table>
<thead>
<tr>
<th>$A-0$</th>
<th>$A-1$</th>
<th>$A-2$</th>
<th>$A-3$</th>
<th>$A-4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A-0$</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>$A-1$</td>
<td>6</td>
<td>11</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>$A-2$</td>
<td>7</td>
<td>8</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>$A-3$</td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>19</td>
</tr>
<tr>
<td>$A-4$</td>
<td>21</td>
<td>22</td>
<td>23</td>
<td>24</td>
</tr>
</tbody>
</table>

Table 3 Each case

Case 1: Regions $A-0$ and $A-0$
From Tables 1 and 2, we see $\Phi = \{\phi\}$ and therefore the system is apparently stable. We have $N(n) = 2^n$.

Case 2: Regions $A-0$ and $A-1$
The system is the place $\Phi = \{(1, -1, 1)\}$. Then equilibrium points are all points except combinations including triple(1, -1, 1) in k-length vector. $a_{n-}^-$ means the code of k-length vector added to $-$ at the last part. Similarly $a_{n+}^+$ is number in the case of vector adding $+$ at the end of code not including (1, -1, 1).

In similar vein with $a_{n-}^+$ and $a_{n+}^+$.

The forbidden code is only $(+ - +)$, we see that the following equations are satisfied:

$$a_{k+1} = a_k^+ + a_k^-$$
$$a_{k+1}^+ = a_k^+ + a_k^-$$
$$a_{k+1}^- = a_k^+ + a_k^-$$

where $a_2^- = a_2^+ = a_2 = a_2^- = 1$. Then we have:

$$a_{k+1} = a_k^+ + a_{k+1}$$

Let $u_k \equiv a_k^-$, $v_k \equiv a_k^-$ and $w_k = a_k^+$ for simply. Then

$$\begin{bmatrix}
  u_{k+1} \\
v_{k+1} \\
w_{k+1}
\end{bmatrix} = \begin{bmatrix}
  1 & 1 & 0 \\
  0 & 0 & 1 \\
  1 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  u_k \\
v_k \\
w_k
\end{bmatrix}$$

from which we have:

$$\begin{bmatrix}
  u_k \\
v_k \\
w_k
\end{bmatrix} = \begin{bmatrix}
  1 & 1 & 0 \\
  0 & 0 & 1 \\
  1 & 1 & 1
\end{bmatrix}$$

By summing total, one immediately obtains the number of equilibrium points not including $(-1, 1, 1)$.

$$N(k) = 2u_k + v_k + w_k \quad (19)$$

By solving the characteristic equation:

$$\begin{vmatrix}
  1 - \lambda & 0 & 1 \\
  0 & -\lambda & 1 \\
  1 & 0 & 1 - \lambda
\end{vmatrix} = -\lambda^3 + 2\lambda^2 - \lambda + 1 = 0 \quad (20)$$

we get the values of zeros as $\lambda_1 \approx 1.75, \lambda_2 \approx 0.12 + 0.12i, \lambda_3 \approx 0.12 - 0.12i$. Consequently we obtain:

$$N(n) \approx 3.98(1.75)^n + 0.97(0.12 + 0.12i)^n + 0.97(0.12 - 0.12i)^n \quad (21)$$

Case 3: Regions $A-0$ and $\bar{A}-2$
Since $\Phi = \{(1, -1, 1), (-1, 1, -1), (-1, -1, -1), 1, 1, 1\}$ from Tables 1 and 2, $S$ is stable. When there is $-1$ in the state vector and the value of left neighbor by the $-1$ is 1, the center value $-1$ changes to 1 in either case of 1 or 0 of right neighbor. consequently, the types as like $(--- \ldots + + + \ldots ++)$ remain as equilibrium points. $N(n; -) = N(n; +) = n, N(n; + -) = N(n; ++) = 1, N(n) = 2n + 2$.

Case 4: Regions $A-0$ and $A-3$
In this case we have $\Phi = \{(-1, 1, -1), (1, -1, -1), (-1, -1, 1)\}$ from Tables 1 and 2 and $S$ is stable. As there is $(1, -1, 1)$, $-1$ changes from left side into all 1's $(+ \ldots + + \ldots +)$ in an avalanche in the boundary condition $x_0(0) = 1$, that is, $N(n; +) = N(n; ++) = 1$. The case $(1, -1, 1)$ similarly yields $N(n; -) = 1$ by changing from inverse direction. For the boundary condition $x_0(0) = x_{n+1}(0) = -1, (+ \ldots + + \ldots +)$ $(-- \ldots -- \ldots -)$ including both boundary
conditons. We see \( N(n; - -) = 2 \). After all, we have \( N(n) = 5 \).

**Case 5:** Regions \( A-0 \) and \( \bar{A}-4 \)

Equilibrium point is only \((+++...+...+...)\), i.e., \( N(n) = 1 \).

**Case 6:** Regions \( A-1 \) and \( \bar{A}-0 \) Similar to Case 2.

**Case 7:** Regions \( A-1 \) and \( \bar{A}-1 \)

We get \( \bar{N}(k + 2) = N(k + 1) + N(k) \). This is the typical Fibonacci series.

\[
N(n) = \left( \frac{1 + \sqrt{5}}{2} \right) \left( \frac{1 + \sqrt{5}}{2} \right)^n + \left( 1 - \frac{1 + 2\sqrt{5}}{2} \right) \left( \frac{1 - \sqrt{5}}{2} \right)^n 
\]

(22)

**Case 8:** Regions \( A-1 \) and \( \bar{A}-2 \)

Since the forbidden triples are \((-1,1,-1), (1,-1,1), (1,-1,-1)\), we get the result by recursive equations similar to the one used in Case 2. Therefore the number \( N(n) \) not including the triples \{(-1,1,-1), (1,-1,1), (1,-1,-1)\} is given as:

\[
N(n) = N(n;--)+N(n;+-)+N(n;+-)+N(n;++)
= 1 + 1 + (n - 2) + (n - 1) = 2n + 1
\]

(23)

**Cases 9:** Regions \( A-1 \) and \( \bar{A}-3 \)

This system is not stable in this case.

**Case 10:** Regions \( A-1 \) and \( \bar{A}-4 \) (Unstable (See Remark 1))

**Cases 11 and 12:** Regions \( A-2, \bar{A}-0 \) and \( A-2, \bar{A}-1 \)

These are similar to Case 3 and 8.

**Case 13:** Regions \( A-2 \) and \( \bar{A}-2 \)

\[
N(n) = 1
\]

(24)

**Case 14:** Regions \( A-2 \) and \( \bar{A}-3 \) (unstable)

\[
N(n;--)=N(n;++)=0, N(n;++)=N(n;++)=1
\]

(25)

**Case 15:** Regions \( A-2 \) and \( \bar{A}-4 \) (unstable)

This case is similar to Case 14 and all \( x(0) \) converge to \((+++...+...+...)\) by depending the boundary condition \( x_0(0) = +1 \).

\[
N(n;--)=N(n;++)=0, N(n;++)=N(n;++)=1
\]

(26)

**Cases 16:** Regions \( A-3, \bar{A}-0 \)

This is similar to Case 4.

**Cases 17-20:** (unstable)

**Cases 21:** This is similar to Case 5.

**Cases 22, 23, 24:** (unstable) These are similar to case 10,15,20 respectively.

**Case 25:** Regions \( A-4 \) and \( \bar{A}-4 \) (Unstable; see Remark 1)

After all, we summarize the number equilibrium points as shown in Table 4, where * means unstable regions.

<table>
<thead>
<tr>
<th>( \bar{A}-0 )</th>
<th>( \bar{A}-1 )</th>
<th>( \bar{A}-1 )</th>
<th>( \bar{A}-3 )</th>
<th>( \bar{A}-4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( 2^n, 2^n, 2^n )</td>
<td>Case 2</td>
<td>( n, n, 1 )</td>
<td>( 2n + 3 )</td>
<td>*</td>
</tr>
<tr>
<td>( A-1 )</td>
<td>Case 6</td>
<td>Case 7</td>
<td>( 2n + 3 )</td>
<td>( 1,1,n )</td>
</tr>
<tr>
<td>( A-2 )</td>
<td>( 1,1,n )</td>
<td>( 2n + 3 )</td>
<td>( 1,1,n )</td>
<td>*</td>
</tr>
<tr>
<td>( A-3 )</td>
<td>( 1,1,1,2 )</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>( A-4 )</td>
<td>( 1,1,1,1 )</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

Table 4 Number of equilibrium points

5. Conclusion

We gave the number of equilibrium point in each region for 1-D DBCNN with no input. As seen from Table 4 the number of equilibrium points is fairly small.

**Acknowledgment**

This research was partially supported by Core Research for Evolutional Science and Technology (CREST), and Japan Science and Technology Agency (JST).

**References**


Prediction of Chaos Attractor by using DT-CNN

Masatomo Inoue\textsuperscript{*}, Yuta Watanabe\textsuperscript{1}, Tatsuro Shimada\textsuperscript{3}, Masayuki Yamauchi\textsuperscript{5}, and Mamoru Tanaka\textsuperscript{*}

\textsuperscript{*} Department of Electrical and Electronics Engineering, Sophia University
7-1, Kioi-cho, Chiyoda-ku, Tokyo, 102-8554, Japan
Email: [masatomo, tanaka]@lab.ee.sophia.ac.jp
\textsuperscript{1} Services Technology Group, Fuji Xerox Co., Ltd. 2274, Honigo, Ebina-shi, Kanagawa, 243-0494, Japan
Email: youta0117@hotmail.com
\textsuperscript{3} Powertrain Development Group, Toyota Motor Corporation
1, Toyota-Cho, Toyota-shi, Aichi, 471-8572, Japan
Email: tatsuro@shimada.tec.toyota.co.jp
\textsuperscript{5} Department of Electronics and Photonic Systems Engineering, Hiroshima Institute of Technology
2-1-1, Miyake, Saeki-ku, Hiroshima, 731-5193, Japan
Email: yamauchi@ieee.org

Abstract—We propose the prediction of time series by using Discrete Time Cellular Neural Network (DT-CNN) learning machine of the differential equations. The coefficients of the equations can be obtained by equilibrium points of the DT-CNN state equation. In this process, Householder Transformation is used to transform the feedback A-template into tri-diagonal sparse matrix.

As a simulation, we predict the chaos attractor of Chua’s circuit. We consider how much the attractor is predicted from the observed time sequence based on our method.

1. Introduction

There is something that changes at time in the world where we live. The majority have the nonlinearity. For example, weather and movement of stock prices are typical. The nonlinear time series analysis and prediction is the important technique in fields such as economics, weather, and engineering systems. In those fields, many kinds of chaos are often observed. Therefore, analysis and prediction of chaos are very important. Many researches have been carried up to now[1][2].

In this paper, we propose the prediction of time series from Discrete Time Cellular Neural Network (DT-CNN)[3]-[5]. The coefficients of the predicted nonlinear differential equations can be obtained by equilibrium points of the DT-CNN state equation. That is, the DT-CNN is used as a machine learning solver. The feedback A- and feed-forward B-templates of its nonlinear state equations are expressed by using past values in observed time sequences. The predicted coefficient matrices corresponding to the templates must be obtained by a new machine learning algorithm. In this process, Householder Transformation is used to transform its dense matrix into the tri-diagonal sparse matrix. The tri-diagonal matrix is corresponding to the feedback A-template of state equation of the DT-CNN. The predicted differential equations are solved by the DT-CNN.

As a simulation, we predict the chaos attractor of Chua’s circuit. We consider how much the attractor is predicted from the observed time sequence based on our method.

2. Machine Learning

Let $u_t$ be an integral continuous variable for observed input variable $du/dt$, then the proposed model is described by using state equation as

$$\frac{du}{dt} = -a_0 u_t + a_1 u_{t-1} + \cdots + a_{ij} u_{t-j} + \cdots + a_N u_{t-N}$$

(1)

where $N$ is the number of input variables, $a_{ij}$ is a coefficient weight connected from cell $j$ to $i$.

From the definition of the derivative, $du/dt$ is expressed as

$$\frac{du}{dt} = \lim_{\Delta t \to 0} \frac{u(t + \Delta t) - u(t)}{\Delta t}.$$  (2)

In the case of $\Delta t = 1$, equation (2) is made discretization as follows:

$$\frac{du}{dt} = u(k + 1) - u(k).$$  (3)

Assuming that there is no rapid changes of $u_t$ from discrete time $k$ to $k + 1$ in the discrete time space, the variable $u_t$ is expressed in the mean value of $u_t(k)$ and $u_t(k + 1)$ as

$$u_t = \frac{1}{2} [u_t(k) + u_t(k + 1)].$$  (4)

Now, in order to use the DT-CNN as a machine learning solver, it is determined that the parameter $a_{ij}$ is corresponding to the nonlinear output $d(x_{ij})$ for state variable $x$.
as \( a_{ij} = d(x_i) \). The nonlinear function \( d(x) \) is a quantizing function to transform the steady state variable \( x \) to the quantized variable for digitalization.

Then, the equation (1) is described by using equations (2), (3), and (4) from discrete time 1 to \( n \) as follows:

\[
\begin{bmatrix}
  u_t(2) - u_t(1) \\
  u_t(3) - u_t(2) \\
  \vdots \\
  u_t(n) - u_t(n-1)
\end{bmatrix} = -d(x_t) = \begin{bmatrix}
  \frac{1}{2}(u_t(1) + u_t(2)) \\
  \frac{1}{2}(u_t(2) + u_t(3)) \\
  \vdots \\
  \frac{1}{2}(u_t(n-1) + u_t(n))
\end{bmatrix}
\]

\[ + d(x_t) \begin{bmatrix}
  u_t(1) \\
  u_t(3) \\
  \vdots \\
  u_t(n)
\end{bmatrix} \begin{bmatrix}
  u_t(2) \\
  u_t(3) \\
  \vdots \\
  u_t(n)
\end{bmatrix} \end{bmatrix} + \cdots + d(x_{t(N-1)}) \begin{bmatrix}
  u_{t(N-1)} \\
  u_{t(N-1)} \\
  \vdots \\
  u_{t(N-1)}
\end{bmatrix} = 0. \tag{5}
\]

Let \( \mathbf{u} = [u_t(2) - u_t(1), u_t(3) - u_t(2), \ldots, u_t(n) - u_t(n-1)]^T \) be a vector for the number \( n-1 \) of past observed samples in time series, then we can derive the following nonlinear algebra equation from equation (5):

\[
\mathbf{u} = \mathbf{GD}(x) \tag{6}
\]

where

\[
\mathbf{D}(x) = \begin{bmatrix}
  d(x_1), d(x_2), \ldots, d(x_N)
\end{bmatrix}^T
\]

and \( \mathbf{G} \) is \((n-1) \times N\) rectangular matrix given by

\[
\mathbf{G} = \begin{bmatrix}
  (-\frac{1}{2}(u_t(1) + u_t(2))) & u_t(2) & u_t(3) & \cdots & u_t(n) \\
  (-\frac{1}{2}(u_t(2) + u_t(3))) & u_t(3) & u_t(4) & \cdots & u_t(n) \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  (-\frac{1}{2}(u_t(n-1) + u_t(n))) & u_t(n) & \cdots & \cdots & u_t(n)
\end{bmatrix}
\]

The transformation to non-singular equation is applied to minimize the norm \( \| \mathbf{GD}(x) - \mathbf{u} \| \) as

\[
\mathbf{G}^T \mathbf{GD}(x) - \mathbf{G}^T \mathbf{u} = 0 \tag{7}
\]

where \( \mathbf{G}^T \) is the transposed matrix of \( \mathbf{G} \).

3. Parameter determination and prediction

Since the matrix \( \mathbf{G}^T \mathbf{G} \) is non-singular, the equation (7) can be solved by the Newton Raphson Method directly if the initial value is in the neighbor of the solution. However, it is very difficult to find the initial value of the Newton Raphson Method. Therefore, we think that we solve the equation (7) by using the state equation of the DT-CNN. That is, the nonlinear coefficient vector \( \mathbf{D}(x) \) can be obtained by

\[
\mathbf{x}(k + 1) = \mathbf{A}\mathbf{x}(k) + \mathbf{Bu} + \mathbf{T} \tag{8}
\]

where

\[
\mathbf{A} = -\mathbf{G}^T \mathbf{G}, \quad \mathbf{B} = \mathbf{G}^T, \quad \mathbf{T} = 0.
\]

For the correspondence to the state equation of the DT-CNN, \( \mathbf{x}(k) \) is a state variable vector, \( \mathbf{u} \) is an input variable vector, \( \mathbf{f}(\mathbf{x}(k)) \) is an output variable vector, \( \mathbf{A} \) and \( \mathbf{B} \) are feed-back and feed-forward weight matrices corresponding to \( \mathbf{A} \)- and \( \mathbf{B} \)-templates respectively. \( \mathbf{f}(\mathbf{x}(k)) \) is assumed to be \( \mathbf{x}(k) \) in linear DT-CNN. Then, we consider the Householder Transformation in order that it is easy to change the dense matrix to the sparse matrix.

The matrix \( \mathbf{A} \) transform into tri-diagonal matrix by the Householder matrix \( \mathbf{P} \). Tri-diagonal matrix \( \mathbf{A}_{3d} \) is expressed as

\[
\mathbf{A}_{3d} = \mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \begin{bmatrix}
  \alpha_1 & \beta_1 & 0 \\
  \beta_1 & \alpha_2 & \beta_2 \\
  \vdots & \vdots & \vdots \\
  0 & \beta_{n-1} & \alpha_n
\end{bmatrix}. \tag{9}
\]

Since \( \mathbf{x}(k) = \mathbf{P}^{-1} \mathbf{y}(k) \) and \( \mathbf{PP}^{-1} = 1 \), the equation (8) can be rewritten as

\[
\mathbf{y}(k + 1) = \mathbf{A}_{3d} \mathbf{y}(k) + \mathbf{PBu}. \tag{10}
\]

Because the \( \mathbf{A} \)-matrix is transformed into the tri-diagonal sparse matrix \( \mathbf{A}_{3d} \) in equation (10), the equation (10) can be rewritten by using the state equation of the DT-CNN for the constant vector \( \mathbf{PBu} \). It makes it possible to reduce the computation drastically. We can prove the convergence of the solution in the equation (10). Basically, the coefficients of the predicted nonlinear differential equation (1) can be obtained by equilibrium points of the DT-CNN state equation (10).

The equations (1) is rewriten as follows:

\[
\begin{bmatrix}
  \frac{d}{dt} u_1 \\
  \frac{d}{dt} u_2 \\
  \vdots \\
  \frac{d}{dt} u_N
\end{bmatrix}^T = 
\begin{bmatrix}
  -d(x_{11}) & d(x_{12}) & \cdots & d(x_{1N}) \\
  d(x_{21}) & -d(x_{22}) & \cdots & d(x_{2N}) \\
  \vdots & \vdots & \ddots & \vdots \\
  d(x_{N1}) & \cdots & -d(x_{NN})
\end{bmatrix} \begin{bmatrix}
  u_1 \\
  u_2 \\
  \vdots \\
  u_N
\end{bmatrix} \tag{11}
\]

Then, the equation (11) is described in discretization as
\[
\begin{bmatrix}
    u_1(k+1) & u_2(k+1) & \cdots & u_N(k+1)
\end{bmatrix}^T = \\
\begin{bmatrix}
    1 - d(x_{11}) & d(x_{12}) & \cdots & d(x_{1N}) \\
    d(x_{21}) & 1 - d(x_{22}) & \cdots & d(x_{2N}) \\
    \vdots & \vdots & \ddots & \vdots \\
    d(x_{N1}) & \cdots & \cdots & 1 - d(x_{NN})
\end{bmatrix} \\
\begin{bmatrix}
    u_1(k) \\
    u_2(k) \\
    \vdots \\
    u_N(k)
\end{bmatrix}
\] (12)

For the DT-CNN representation, the matrices are given by

\[
A = \begin{bmatrix}
    1 - d(x_{11}) & d(x_{12}) & \cdots & d(x_{1N}) \\
    d(x_{21}) & 1 - d(x_{22}) & \cdots & d(x_{2N}) \\
    \vdots & \vdots & \ddots & \vdots \\
    d(x_{N1}) & \cdots & \cdots & 1 - d(x_{NN})
\end{bmatrix}
\]

and \(B = T = 0\).

The flowchart of our proposed method is shown in the Fig.1.

4. Experiment

We predict the chaos attractor with this algorithm. The attractor that we use is Chua’s attractor.

4.1. Chua’s circuit

Chua’s circuit[8] is shown in Fig.2. It consists of two capacitors, one inductor, one linear resistor, and one non-linear resistor. The \(v - i\) characteristic of the non-linear resistor is shown in Fig.3. It is described by following set of normalized differential equations

\[
\begin{align*}
\frac{dx}{dt} &= \alpha (y - x - f(x)) \\
\frac{dy}{dt} &= x - y + z \\
\frac{dz}{dt} &= -\beta y
\end{align*}
\]

(13)

\[
f(x) = bx + \frac{1}{2} (a - b) \left| x + 1 \right| - \left| x - 1 \right|
\]

(14)

where

\[
x = \frac{V_1}{B_p}, \quad y = \frac{V_2}{B_p}, \quad z = i_3 \frac{R}{B_p},
\]

\[
\alpha = \frac{C_2}{C_1}, \quad \beta = \frac{R^2 C_2}{L}, \quad a = RG_a, \text{ and } b = RG_b.
\]

Figure 2: Chua’s circuit.

Figure 3: The \(v - i\) characteristic of the non-linear resistor.

The attractor of the Fig.4 can be generated by using \(\alpha = 8.99, \beta = 12.35, a = -1.09\), and \(b = -0.57\) in the equations(13) by Runge-Kutta Method.

4.2. Simulation result

The comparison between original Chua’s attractor and predicted one from 100 datasets to 1 dataset by using the DT-CNN is shown in the Fig.5. \(N\) in the equation(1) is 3...
because Chua’s equations are three dimensions. The predicted attractor is almost same as the original attractor derived from the differential equations of the Chua’s circuit (Fig. 4).

The Fig. 6 is the root-mean-square error (RMSE) of the attractor which is predicted by using the DT-CNN when the predicted time series dataset is 1 and time series datasets used to the prediction is changed. When time series datasets is increased, the RMSE is decreased.

5. Conclusion

We could predict time series by using DT-CNN. The attractor which is predicted from the observed time sequence based on our method was almost same as that of attractor derived from the differential equation of the Chua’s Circuit. The error was decreased when time series datasets was increased. In the future, we will predict other chaos attractor and investigate the error. And we’ll predict time series by using the predicted value.

Acknowledgment

This research is supported by the fund of Open Research Center Project from MEXT of Japanese Government (2002-2006).
Further Analysis on Condition for 1-D CNNs to Perform Connected Component Detection

Norikazu Takahashi † and Tetsuo Nishi ‡

† Department of Computer Science and Communication Engineering, Kyushu University
6-10-1 Hakozaki, Higashi-ku, Fukuoka 812-8581 Japan
‡ Faculty of Science and Engineering, Waseda University
Tokyo 162-0072 Japan
Email: norikazu@csce.kyushu-u.ac.jp, nishi-t@waseda.jp

Abstract—The authors have recently derived a sufficient condition for one-dimensional cellular neural networks to perform connected component detection [1]. However, this condition is expressed as a very complex inequality and therefore it is not so easy to check whether given parameters satisfy the condition or not. In this paper, we will analyze the inequality in detail and present much simpler sufficient conditions.

1. Introduction

Let us consider one-dimensional cellular neural networks (1-D CNNs) with the antisymmetric template \( A = [s, p, -s] \). The behavior of such a CNN is described by the following differential equations:

\[
\dot{x}_i = -x_i + s y_{i-1} + p y_i - s y_{i+1}, \quad i = 1, 2, \ldots, n
\]

(1)

where \( x_i \) and \( y_i \) represent the state and output of the \( i \)-th cell, respectively. Relationship between the state \( x_i \) and the output \( y_i \) is expressed with a piecewise-linear function as follows:

\[
y_i = f(x_i) \triangleq \frac{1}{2} y(|x_i + 1| - |x_i - 1|).
\]

(2)

An important application of these 1-D CNNs is connected component detection (CCD). Given a 1-D black-and-white image, CCD is the task to obtain the output image satisfying 1) the number of black pixels is equal to the number of blocks of consecutive black pixels in the input image, 2) the right-most pixel is black unless all pixels are white, 3) all black pixels are isolated, and 4) there exists exactly one white pixel between two neighboring black pixels (see Fig.1). Matsumoto et al. [2] showed experimentally that if a 1-D CNN has the template \( A = [1, 2, -1] \) and the initial state is black-and-white, that is, \( x_i(0) \in [-1, 1], \forall i \) then the CNN can perform CCD. Since then the global dynamics of 1-D CNNs with the antisymmetric template has been vigorously studied [3]–[8]. Zou et al. [3] proved, under the boundary condition: \( y_0(t) = y_{n+1}(t) = -1, \forall t \)

(3)

and shown that such a CNN can perform CCD if the template \( A = [s, p, -s] \) satisfies

\[
s > (p - 1)/2 > 0
\]

(4)

and

\[
g(p, s) \triangleq (2s + p - 1)^{1/2} [p(p - 1) + s(2s - p - 1)]
\]

\[-(2s + p + 1)^{1/2} (p(p - 1) + s(2s - p - 1)) \geq 0.
\]

(5)

Figure 1 shows two curves expressed by \( s = p - 1 \) and \( g(p, s) = 0 \). Since the equation \( g(p, s) = 0 \) cannot be solved...
for \( s \) analytically, we have solved it numerically by using the bisection method for each value of \( p \). If we take a look at Fig.2, we can easily expect that for each value of \( p \) the equation \( g(p, s) = 0 \) has a unique solution and therefore for any point in the region bounded by the two curves satisfies (4) and (5). However, this expectation has never been proved analytically so far.

![Figure 2: Two curves expressed by \( s = (p - 1)/2 \) and \( g(p, s) = 0 \).](image)

The purpose of this paper is to study some basic properties of the function \( g(p, s) \) analytically and to derive simpler conditions for 1-D CNNs described by (1) to be able to perform CCD.

2. Analysis of Function \( g(p, s) \)

**Proposition 1** If \( s = (p - 1)/2 \) then \( g(p, s) > 0 \).

**Proof:** Substituting \( s = (p - 1)/2 \) into \( g(p, s) \), we have

\[
g(p, (p - 1)/2) = \left(2(p - 1)\right)^{1/2} \left(p(p - 1) + \frac{p - 1}{2}(p - 1 - p - 1)\right)
\]

\[
= (2(p - 1))^{1/2}(p - 1)^2
\]

which is positive for all \( p > 1 \). \( \square \)

**Proposition 2** If \( p = 2 \) then \( g(p, s) \) satisfies

\[
g(p, s) = \begin{cases} 
> 0 & \text{if } s < 1 \\
= 0 & \text{if } s = 1 \\
< 0 & \text{if } s > 1
\end{cases}
\]

**Proof:** Substituting \( p = 2 \) into \( g(p, s) \), we have

\[
g(2, s)
= (2s + 1)[2 + s(2s - 3)] - (2s - 1)[2 + s(2s - 1)]
= (2s + 1)(2s^2 - 3s + 2) - (2s - 1)(2s^2 - s + 2)
= -4s^2 + 4s^2 - 4s + 4
= 4(1 - s)
\]

which completes the proof. \( \square \)

**Proposition 3** If \( p > 2 \) and \( s = p - 1 \) then \( g(p, s) > 0 \).

**Proof:** Substituting \( s = p - 1 \) into \( g(p, s) \), we have

\[
g(p, p - 1)
\]

\[
= \left(2(p - 1)\right)^{1/2}\left(p(p - 1) + (p - 1)(2p - 2 - p - 1)\right)
\]

\[
-(p - 1)^{1/2}(p(p - 1) + (p - 1)(2p - 2 + p + 1))
\]

\[
= (p - 1)^{1/2}\left(3^{1/2}(2p - 3) - (2p - 1)\right)
\]

Since \( (p - 1)^{1/2} > 0 \) for all \( p > 1 \), it suffices for us to show that the function \( g_1(p) \) defined by

\[
g_1(p) \pm 3^{1/2}(2p - 3) - (2p - 1)
\]

is positive for all \( p > 2 \). Let us first find the value of \( g_1(p) \) at \( p = 2 \) and \( p = \infty \). Substituting \( p = 2 \) into the right-hand side of (7), we have

\[
g_1(2) = 3 \cdot 1 - 3 = 0.
\]

Letting \( p \) go to infinity in the right-hand side of (7), we have

\[
\lim_{p \to \infty} g_1(p) = 2 \lim_{p \to \infty} p(3^{1/2} - 1) - 2.
\]

By applying the De L’Hôpital’s theorem to the first term, we have

\[
\lim_{p \to \infty} p(3^{1/2} - 1) = \lim_{p \to \infty} \frac{3^{1/2} - 1}{1/p}
\]

\[
= \lim_{p \to \infty} \frac{(3^{1/2} - 1)'}{(1/p)'}
\]

\[
= \lim_{p \to \infty} (\ln 3) \cdot \frac{p^2}{(p - 1)^2} \cdot 3^{1/2}
\]

\[
= \ln 3.
\]

Therefore we have

\[
\lim_{p \to \infty} g_1(p) = 2 \ln 3 - 2 \approx 0.197225 \cdots > 0.
\]

Next, in order to understand the properties of the function \( g_1(p) \) further, let us consider the first and second derivatives. The first derivative of \( g_1(p) \) is given by

\[
\frac{dg_1(p)}{dp} = -(\ln 3) \cdot \frac{3^{1/2}}{(p - 1)^2} \cdot (2p - 3) + 2 \cdot 3^{1/2} - 2
\]

\[
= 3^{1/2}\left(2 - (\ln 3)(2p - 3)\right) - 2.
\]

Substituting \( p = 2 \) into the right-hand side, we have

\[
\frac{dg_1(p)}{dp} \bigg|_{p=2} = 3(2 - \ln 3) - 2 > 0.
\]

Also, by letting \( p \) go to infinity in the right-hand side, we have

\[
\lim_{p \to \infty} \frac{dg_1(p)}{dp} = 0.
\]
The second derivative of \( g_1(p) \) is given by
\[
\frac{d^2 g_1(p)}{dp^2} = -\ln 3 \cdot \frac{3^{\frac{1}{p}}}{(p-1)^2} \cdot \left\{ 2 - \ln 3 \cdot \frac{(2p-3)}{(p-1)^2} \right\} + 3^{\frac{1}{p}} \left\{ -\ln 3 \cdot \frac{2(p-1)^2 - (2p-3) \cdot 2 \cdot (p-1)}{(p-1)^4} \right\} = \frac{\ln 3 \cdot 3^{\frac{1}{p}}}{(p-1)^4} \left\{ \ln 3 \cdot (2p-3) - 2(p-1) \right\}.
\]
from which we have
\[
\frac{d^2 g_1(p)}{dp^2} \begin{cases} < 0, & \text{if } p < p^* \\ = 0, & \text{if } p = p^* \\ > 0, & \text{if } p > p^* \end{cases} \quad (12)
\]
where \( p^* \) is given by
\[
p^* = \frac{3 \ln 3 - 2}{2(\ln 3 - 1)} \approx 6.57.
\]
Eq.(12) means that the first derivative \( \frac{dg_1(p)}{dp} \) is strictly monotone decreasing for \( 1 < p < p^* \) and strictly monotone increasing for \( p > p^* \). From this fact together with (11), we can easily show that
\[
\left. \frac{dg_1(p)}{dp} \right|_{p=p^*} < 0 \quad (13)
\]
holds. To do so, let us assume that (13) does not hold. Then, since \( \frac{dg_1(p)}{dp} \) is strictly monotone increasing for \( p > p^* \), \( \lim_{p \to \infty} \frac{dg_1(p)}{dp} \) must be positive. However, this contradicts (11). Furthermore, it follows from (10) and (13) that there exists \( \hat{p} \) in the interval \((2, p^*)\) such that
\[
\frac{dg_1(p)}{dp} \begin{cases} > 0, & \text{if } 2 < p < \hat{p} \\ = 0, & \text{if } p = \hat{p} \\ < 0, & \text{if } \hat{p} < p \end{cases} \quad (14)
\]
holds. Now we are ready for proving the proposition. Since \( g_1(p) \) is strictly monotone decreasing for \( p > p^* \) and approaches 0.197225⋅⋅⋅ as \( p \to \infty \) (see (9) and (14)), we can conclude that \( g_1(p) > 0 \) for \( p \geq p^* \). Moreover, since \( g_1(2) = 0, g_1(p^*) > 0, \) and \( g_1(p) \) is concave for \( 2 \leq p \leq p^* \) (see (12)), we can conclude that \( g_1(p) > 0 \) for all \( p \in (2, p^*) \). This completes the proof. \( \square \)

**Proposition 4** If \( 1 < p < 2 \) and \( s = p - 1 \) then \( g(p, s) < 0 \).

**Proof:** As we have seen in the proof of Proposition 3, it suffices for us to show that \( g_1(p) \) is negative for \( 1 < p < 2 \). Since the first derivative of \( g_1(p) \) is positive at \( p = 2 \) (see (10)) and the second derivative is negative for \( 1 < p < p^* \) (see (12)), \( g_1(p) \) is monotone increasing for \( 1 < p < 2 \). Since \( g_1(2) = 0, g_1(p) \) is negative for \( 1 < p < 2 \). \( \square \)

Figure 3 shows the graphs of \( g_1(p) \) and its first and second derivatives.

![Figure 3: Graphs of \( g_1(p) \) (upper) and its derivatives (lower).](image)

**Proposition 5** Let \( s = a(p-1) \) where \( a > 1/2 \). Then \( \lim_{p \to \infty} g(p, s) \geq 0 \) if and only if
\[
h(a) = \ln \left( \frac{2a + 1}{2a - 1} \right) \cdot (2a^2 - a + 1) - 2a \geq 0 \quad (15)
\]

**Proof:** Substituting \( s = a(p-1) \) into \( g(p, s) \), we have
\[
g(p, a(p-1)) = \left\{ (2a+1)(p-1) \right\}^{\frac{1}{2a-1}} \cdot (2a^2 - a + 1) - 2a \geq 0
\]
where
\[
g_2(p) = (2a^2 - a + 1) \left\{ (2a+1)^{\frac{1}{2a}} - (2a-1)^{\frac{1}{2a}} \right\} p
\]
\[
- a \left\{ (2a+1)^{\frac{1}{2a}} - (2a-1)^{\frac{1}{2a}} \right\}.
\]
By using the De L’Hôpital’s theorem, we have
\[
\lim_{p \to \infty} (2a+1)^{\frac{1}{2a}} = \lim_{p \to \infty} \frac{(2a+1)^{\frac{1}{2a}}}{1/p} = \lim_{p \to \infty} -\ln(2a+1) \cdot (2a+1)^{\frac{1}{2a}} / (p-1)^2 = \ln(2a+1).
\]
Similarly we have
\[ \lim_{p \to \infty} (2a - 1)^{\frac{1}{p}} = \ln(2a - 1). \] (17)

Letting \( p \) go to infinity in \( g_s(p) \) and substituting (16) and (17), we have
\[ \lim_{p \to \infty} g_s(p) = (2a^2 - a + 1)(\ln(2a + 1) - \ln(2a - 1)) - 2a \]
which is nonnegative if and only if (15) holds.

Figure 4 shows the graph of \( h(a) \). Although we have not proved analytically, it is expected from the figure that \( h(a) \) is monotone decreasing and has a unique zero \( a^* \). By solving the equation \( h(a) = 0 \) numerically, we have
\[ a^* = 1.1755715 \cdots. \]

3. Simpler Sufficient Conditions for CCD

From Propositions 2 and 3, we can obtain simpler sufficient conditions for 1-D CNNs to perform CCD as follows:

**Theorem 1** A 1-D CNN described by (1)–(4) can perform CCD if one of the following conditions is satisfied.

1) \( p = 2 \) and \( 0.5 < s \leq 1 \).
2) \( p > 2 \) and \( s = p - 1 \).

Also, it follows from Proposition 1 that if \( s - (p - 2)/2 \) is sufficiently small then the 1-D CNN can perform CCD. Furthermore, it follows from Proposition 5 that if \( p \) is sufficiently large and \( (p - 2)/2 < s < (p - 1) \times 1.1755715 \cdots \) then the 1-D CNN can perform CCD.

Although it was not proved in this paper, the authors expect that \( g(p, s) \) has no zero in the region defined by \( \{ (p, s) | p \geq 2, (p - 1)/2 < s \leq p - 1 \} \). This expectation has already been verified numerically, as shown in Fig.1. We finally present this expectation as a conjecture.

**Conjecture 1** A 1-D CNN described by (1)–(4) can perform CCD if the template satisfies \( p \geq 2 \) and \( (p - 1)/2 < s \leq p - 1 \).

4. Concluding Remarks

We have clarified some important properties of the function \( g(p, s) \), and derived much simpler sufficient conditions for 1-D CNNs to perform CCD. However, there still exist some problems to solve. In particular, to make clear the validity of Conjecture 1 is one of the future problems. Also, since we have considered so far only the case where the initial image is binary, to extend the results to the case of non-binary initial image is another future problem.

Acknowledgments

This research was partly supported by the 21st Century COE Program ‘Reconstruction of Social Infrastructure Related to Information Science and Electrical Engineering’.

References

Abstract—In this paper a second order nonlinear system which is inspired by a VLSI oriented continuous-time cellular neural network cell model has been studied. By taking a one-dimensional CNN consisting of the system, travelling waves are obtained. The results have been illustrated with computer simulations using raster plot.

1. Introduction

Over the past few decades, spatio-temporal pattern formation has been actively studied both theoretically and experimentally [5, 12, 3]. As a result of these active studies, today spatio-temporal dynamics are became a tool to solve complex engineering problems such as real-time robot navigation [1] and image processing problems[7].

The Cellular Neural Networks (CNNs) framework provides a very useful methodology for studying spatio-temporal pattern formation [2, 10, 4, 9, 6]. Recently, spatial-temporal phenomena on CNNs have been experimentally verified on the CACE1k (Complex Analogic Cellular Engine) chip [3] which is a VLSI implementation of a two layer CNN with Full-range CNN model. This prototype chip contains 32 × 32 identical cells and it is a two layers Analogic Cellular Engine (ACE) chips which are designed based on CNN-Universal Machine (CNN-UM) architecture [11]. More recently experimental observations of spatio-temporal pattern phenomena on an ACE16k (which has an array of 128 × 128 identical cells on a single layer) chip has been reported by Yalcın et al. in [12] and [13].

In this work, we will study a second order cell model which is inspired by analog Full-range CNN cell. First order of full-range CNN cell model is used in VLSI implementation of CNN. New cell model can be thought as a square wave generator. Here we examine the behaviour of a one-dimensional CNN consisting of identical square wave generator cells with diffusive coupling. Simulations show that this network can generate travelling waves and characteristic of waves can be changed with coupling weights. Furthermore simulations show that the network can be synchronized and stabilized.

This paper is organized as follows. Section 2 briefly describes the new cell model. Section 3 presents pattern formation on a one-dimensional CNN consisting of the new model.

2. A new cell model: square wave generator

The cell dynamic of VLSI oriented continuous-time CNN [8] can be described by the following mathematical model

\[
\dot{x} = -g(x) + I
\]

where \( x \) state variable, \( g(x) \) is the nonlinearity of the cell which is given in Figure 1 and \( I \) is the synaptic law [4]. The nonlinearity can be implemented using two diodes and a reference level [8]. The synaptic law defines the coupling between the cells. Here we consider a second-order cell based on model (1). The cell dynamic of an uncoupled cell is described by the state equations

\[
\begin{align*}
\dot{x} &= ax + by - g(x) \\
\dot{y} &= cx + dy
\end{align*}
\]

where \( a, b, c, d \in R \) and \( x, y \in R \) are state variables.

Figure 1: The nonlinear function for \( g(x) \).

Figure 2 shows the limit cycle of the system (2) using the parameters are \( a = 3, b = -4, c = 0.1, d = -0.1, \) and \( m = 100 \). It does appear qualitatively square wave. The temporal evolution of both variables are shown in Figure 3.
3. 1-D array of square wave generator cells

In this work we present spatio-temporal dynamics of locally coupled 128 square wave generator cells. The cells dynamics are described by

\[
\begin{align*}
\dot{x}_i &= ax_i + by_i - g(x_i) + \lambda_0 x_i + \lambda_+ x_{i+1} \\
\dot{y}_i &= cx_i + dy_i
\end{align*}
\]

where \(i = 1, 2, ..., 128\) denotes the index of the cell. Each cell is connected with the neighbour cells. We impose the boundary conditions \(x_0 = x_{128}\) and \(x_{129} = x_1\). In order to observe spatio-temporal dynamics behaviour we use a raster plot. When the \(x\) state variable value of a cell is greater or equal than 1 the cell is coded with black dot in the raster plot. When the \(x\) state variable value of a cell is less or equal than 1 the cell is coded with white dot. Between the 1 and \(\square\) the cell is coded with a gray level. Figure 4 shows how the cell is shown in raster plot for a given temporal evolution of \(x\) state variable.

Here the equations (3) were simulated using Runge Kutta 4 order of integration method taking a time step equal to 0.01. Raster plots of the system (3) are demonstrated spatio-temporal sequences between the iteration steps 500000 and 505500. Figure 5 shows the raster plot of uncoupled 128 cells of the system (2) (equal to the system (3) with \(\lambda_- = 0, \lambda_+ = 0\) and \(\lambda_0 = 0\)). We start to present stable patterns which are result of stability of the cells. For \(\lambda_0 = 0\) and \(\lambda_- \leq 0.5\) one can obtained stable patterns. Figure 6(a) shows a stable pattern obtained for \(\lambda_- = -1\), \(\lambda_+ = 0\) and \(\lambda_0 = 0\). These type of patterns do not include any temporal change. Figure 6(b) shows a raster plot of the system 3 with \(\lambda_- = 1, \lambda_+ = 1, \lambda_0 = -2\). This pattern is a result of complete synchronization [13].

Figure 7 shows the obtained travelling waves. The direction of propagation of waves can be controlled by the coupling weight (see Figure 7 (a) and (b)) and also width of the waves can be adjusted by the coupling weight. The direction of propagation of the waves in Figures 7 (a) and (c) are the same but the wave in Figure 7 (c) is wider than Figure 7 (a) which is obtained by increasing only the value of \(\lambda_-\).

The most interestingly, the array for \(\lambda_0 = 0, \lambda_- = .1\) and \(\lambda_+ = .1\) generates travelling waves which are propagated from the corners of the array. The waves in Figure 8 show the fundamental properties of autowaves: two waves that are spreading in opposite directions do not pass each other but mutually annihilate.

A chessboard-type pattern has been also obtained form the array. Figure 9 shows the obtained patterns. Time waveforms of the first four state variables \((x_i, i = 1, 2, 3, 4)\) of the array are given in Figure 10. Temporal evolution of the pattern can be controlled by the coupling weight. When
Figure 6: Raster plots of the system 3 with coupling weights (a) $\lambda = □1, \lambda = 0, \lambda_+ = □1$ and (b) $\lambda = 1, \lambda_+ = \lambda_0 = □2$ shows a stable pattern and a pattern which is a result of synchronization, respectively.

the coupling weight $\lambda_-$ is decreased, frequency of temporal behaviour is increased (see Figure 9(b)).

4. Conclusions

In this work we have studied the behaviour of a one-dimensional CNN consisting of identical 128 square wave generator cells with diffusive coupling. Travelling waves has been obtained on the network. Furthermore we have observed complete synchronization on the network.

Acknowledgments

This work was supported by The Scientific and Technical Research Council of Turkey, under project 105E103.

References


Figure 8: Raster plots of the array for \( \lambda_0 = 0 \) and \( \lambda_0 = 0.1, \lambda_+ = 0.1 \). Travelling waves which are propagated from the corners of the array. Two waves that are spreading in opposite directions do not pass each other but mutually annihilate.


Global Optimization with Coupled Local Minimizers
Excited by Gaussian White Noise

Serkan Günel, Johan A. K. Suykens, and Joos P. L. Vandewalle

Katholieke Universiteit Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10, B-3001 Leuven (Heverlee), Belgium
Email: {serkan.gunel,johan.suykens,joos.vandewalle}@esat.kuleuven.be

Abstract—In this paper, Stochastic Coupled Local Minimizers (SCLMs) are presented for global optimization of smooth cost functions. This extends the deterministic coupled local minimizers (CLMs) whose best performing elements impose their dynamics to others by means of master–slave synchronization. The elements of the CLM network are excited by Gaussian white noise. An adaptive cooling schedule that effectively decouples the noise sources from the network when a solution candidate is agreed among all CLMs is proposed. Examples are given to illustrate the proposed scheme.

1. Introduction

In many optimization problems, global exploration of the search space is required. However, an intelligent way of exploration is needed, since exhaustive searches are usually too expensive for applications of interest. Many popular methods for global optimization incorporate randomness to achieve global exploration of the search space either in assigning new candidate solutions randomly (e.g., simulated annealing algorithms [1], genetic algorithms [2]) or starting from multiple random solutions with a prior distribution of initial candidates. Exhaustive search is avoided by decreasing the randomness gradually (i.e., cooling in simulated annealers, decreasing mutation rate in genetic algorithms), and combining good candidates of solutions such as crossover operations in genetic algorithms. Satisfactory exploration of the search space can be done with processes running in parallel, and the best solutions of the parallel running processes are the most likely candidates of the global optima.

Recently, Suykens et al. proposed a new scheme which depends on coupling of local optimization processes in order to achieve global optimization of smooth functions, and showed that this approach performs better than conventional multi–start local optimizations consisting of independent runs [3, 4]. The optimization is done by considering the augmented Lagrangian in terms of the average ensemble energy of the individual local minimizers, and the hard and soft constraints of the synchronization. This augmented Lagrangian leads to a set of coupled Lagrange programming networks [5] which exchange information via the coupling. The coupling strength and learning rates of the Coupled Local Minimizers (CLM) are adapted to achieve maximal decrease in the cost, which leads to cooperative behavior. The approach has been tested e.g. with the optimization of Lennard-Jones Clusters, and supervised training of neural networks [3], and the CLM scheme is well suited to a large family of continuous optimization problems.

If all CLMs happen to be at the basin of a local minimum initially, the optimizations might end up in the associated local minimum in some cases, resulting in poor exploration of the search space. Previously, chaotic signals have been proposed as inputs to CLMs in order to improve performance in such situations [6]. In this paper, the injection of white Gaussian noise to several elements of the CLM network has been proposed to improve the overall optimization performance.

We also propose a cooling scheme for changing the noise amplitude, by letting it to be constant and proportional to the energy of the costates of the Lagrangian network (i.e., Lagrange multipliers of the augmented Lagrangian) between the epochs of coupling adaptations of the system. The costates contain information about the pairwise synchronization errors. When the network is away from synchronization, the energy of the costate vector is large and the elements excited by Gaussian white noise are free to explore the search space. When the network tends to synchronize, the noise amplitude is decreased when the energy of the costates decreases.

This paper is organized as follows. In Section 2 the deterministic CLM approach to global optimization problem is revisited. The stochastic CLMs excited with Gaussian white noise are presented in the Section 3. In Section 4 the numerical scheme is given followed by two examples to illustrate the usefulness of the proposed scheme.

2. Deterministic Coupled Local Minimizers

2.1. Basic Formulation

The minimization problem of a twice differentiable function, $U : \mathbb{R}^n \to \mathbb{R}$ is recast into following form when $q$ local minimizers with states $x^{(i)} \in \mathbb{R}^n, i = 1, \ldots, q$ are considered [3],

$$\min_{x^{(i)}} \sum_{i=1}^{q} U(x^{(i)}), \text{ subject to } x^{(i)} - x^{(i+1)} = 0, i = 1, \ldots, q$$ (1)

with boundary conditions $x^{(0)} = x^{(q)}, x^{(q+1)} = x^{(1)}$. The associated augmented Lagrangian of the problem (1), with soft and hard constraints for the pairwise synchronization...
of each local minimizer, is then
\[ L(x^{(i)}, \lambda^{(i)}) = \frac{2}{q} \sum_{t=1}^{n} U(x^{(i)}) + \frac{1}{2} \sum_{t=1}^{n} \gamma_t \|x^{(i)} - x^{(i+1)}\|^2 + \sum_{t=1}^{n} (\lambda^{(i)}, x^{(i)} - x^{(i+1)}) \]
where \( x^{(i)} \in \mathbb{R}^q, i = 1, 2, \ldots, q \) are the Lagrange multipliers and \( \gamma_t \) are the weights for soft constraints [3]. The Lagrange programming network [5] for optimization w.r.t. this Lagrangian forms the following CLM network
\[ \dot{x}^{(i)} = -\gamma_t \nabla_{x^{(i)}} U(x^{(i)}) + \gamma_{i-1} (x^{(i-1)} - x^{(i)}) - \gamma_t (A^{(i)} - x^{(i+1)}) + A^{(i)} - A^{(i-1)} - \lambda^{(i)} \]
\[ A^{(i)} = x^{(i)} - x^{(i+1)}, \quad i = 1, 2, \ldots, q \]
where \( \eta > 0 \) is the learning rate.

### 2.2. Optimal Cooperative Behavior

To ensure optimal cooperative behavior of the CLMs, the coupling weights are adapted at each time interval \( k\Delta T \leq t < (k+1)\Delta T, k = 0, 1, \ldots \) by solving the linear program
\[ \min_{\gamma \in \mathbb{R}^q} \sum_{i=1}^{q} \frac{dU(x^{(i)})}{dt} \mid_{t=k\Delta T} \quad \text{subject to} \quad \gamma \leq \gamma_t \leq \bar{\gamma} \]
where \( \gamma = \{\gamma_1, \gamma_2, \ldots, \gamma_q\}^T, \bar{\gamma} \) and \( \bar{\gamma} \) are user specified upper and lower bounds, respectively [3]. The coupling weights are kept constant in between intervals. At each interval, the step size \( \eta \) is determined by
\[ \eta = q \sum_{i=1}^{q} \left( \frac{\partial U(x^{(i)})}{\partial x^{(i)}} \cdot \mathbf{h} \right) + qa \sum_{i=1}^{q} U(x^{(i)}) / \sum_{i=1}^{q} \left( \frac{\partial U(x^{(i)})}{\partial x^{(i)}} \cdot \partial U(x^{(i)}) / \partial x^{(i)} \right) \]
where \( \mathbf{h} \triangleq \gamma_{i-1} (x^{(i-1)} - x^{(i)}) - \gamma_t (x^{(i)} - x^{(i+1)}) + A^{(i-1)} - \lambda^{(i)} \) to impose exponential decrease of the cost function. \( \alpha \) is a positive constant and \( \eta \) is also subject to additional user defined constraints \( \eta \leq \eta \leq \bar{\eta} \). The \( \gamma \) is assumed to be positive everywhere (if not one adds a constant value to the cost to guarantee this). The difference between the coupling constants \( \gamma_i \) results in master–slave type synchronization dynamics, hence the successful local optimizers impose their dynamics on the others.

Note that only pairwise synchronization constraints are considered in (3). The possibility of more complex interactions and use of inequality constraints instead of equality constraints are also discussed in [4]. It has been shown that when the general distribution of the states, the CLMs provide good optimization performance at roughly \( q \) times the computational cost of standard gradient descent based methods.

However, if the initial distribution of the states are such that all states are contained in a basin of a local minimum, the network may converge to the local minimum, without exploring the search space for the other possibilities. This problem can be solved by applying white noise as inputs into some of the local optimizers. In this case, some of the local optimizers are continuous–time stochastic ana\-\n\[ \text{3. Stochastic Coupled Local Minimizers} \]

### 3.1. Excitation of CLMs with Gaussian white noise

The evolution equations for the CLMs with Gaussian white noise proposed in this paper is given by the following set of stochastic differential equations
\[ \frac{dx^{(i)}}{dt} = -\frac{\eta}{q} \nabla_{x^{(i)}} U(x^{(i)}) + \gamma_{i-1} (x^{(i-1)} - x^{(i)}) dt + \gamma_t (x^{(i)} - x^{(i+1)}) dt + (x^{(i)} - A^{(i)}) dt + G(t) \mathbf{d}^{(i)} \]
where \( \mathbf{d}^{(i)} = (\omega_{1}^{(i)}, \omega_{2}^{(i)}, \ldots, \omega_{n}^{(i)}) \) are \( n \)-dimensional Wiener processes, with \( \omega_{j}^{(i)}, k = 1, \ldots, n \) are independent 1–dimensional Wiener processes. Although the Wiener processes are nowhere differentiable, \( \mathbf{d}^{(i)} dt \) can be considered as the white noise for notational purposes.

Initially, the noise amplitude must be large and the dynam-
\[ \text{3.2. Determination of the Noise Coupling Matrix} \]

The combination of continuous–time annealers with the deterministic local minimizers results in better exploration of the search space together with above mentioned advantages of CLMs, if the cooling schedule is selected properly. Initially, the noise amplitude must be large and the dynamics have to be dominated by noise to let the elements of the network be excited by the white Gaussian noise exploring the search space almost independent of other elements. An obvious choice is to use the same cooling schedules as in continuous–time simulated annealing algorithms [9]. However, when the group dynamics tend to a candidate minimum in average, the exploration of the noisy cells has to be canceled. The whole network tends to minimize the average cost and keeping individual CLMs together because of the synchronization constraints. A minimum is achieved when all local minimizers agree (i.e., the constraints are satisfied) and the noise amplitude is zero.

The required information for this is already contained in the costate \( \lambda \). Defining \( u^{(i)}(k) \triangleq \lambda^{(i)} ((k-1)\Delta T) - \lambda^{(i)} (k\Delta T) \)
and $e^{(i)}_j \equiv x_j^{(i)} - x_j^{(i+1)}$, $i = 1, 2, \ldots, q$, $j = 1, 2, \ldots, n$ and $k = 0, 1, \ldots, q$. 

By the mean value theorem one has, 

$$
\int_{t-\Delta T}^{t}\mathbf{e}^{(i)}_j(t) \, dt = \Delta T \mathbf{e}^{(i)}_j(t_j), \quad t_j \in (k-1)\Delta T, k\Delta T.
$$

Hence, 

$$
\frac{1}{qn} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{e}^{(i)}_j(t_j).
$$

If $\Delta T$ is sufficiently small, the right hand side of Eq. (9) can be viewed as the average energy of the synchronization error, which tends to zero if all CLMs agree on a solution candidate. The noise is effectively decoupled from the system when all the local optimizers tend to agree on a solution, by making use of this and specifically selecting the noise coupling matrices constant in the interval as $k\Delta T \leq t < (k + 1)\Delta T$, $k = 0, 1, \ldots$

$$
G^{(i)}_{\psi}(\delta) = \delta_{\psi} T N \sum_{i=1}^{n} \mathbf{e}^{(i)}_j(u^{(i)}(k), u^{(i)}(\Delta k), s \in N \subset \{1, 2, \ldots, n\}.
$$

where $\delta_{\psi}$ is the Kronecker delta, $T > 0$. $G^{(i)}_{\psi}$ represents the noise gain of the $p^{th}$ white noise source, coupled to $p^{th}$ state of the $q^{th}$ stochastic cell and $N$ is an index set of states, with $r, p = 1, \ldots, n, i, j = 1, \ldots, q$.

For complex problems, $T$ can be changed with time to improve global exploration of the search space, e.g. letting $T(k) = \mu^k T_0$, $T_0 > 0$, and $0 < \mu \leq 1$. If the states of the optimizers are not close to each other initially, the noise amplitude is large, hence, even if all optimizers are in the same basin of a local minimum, the system has a chance to explore other possible locations. The adjustments to the coupling weights and the learning rate are made as in the deterministic CLM case at each interval. If the stochastic cells happen to be near an optimum that is better than the deterministic cells the network behavior is attracted towards the behavior of the stochastic cells. An overall effect of cooperation that is taking place between the stochastic cells that are trying to explore the space and the deterministic cells that are trying to keep synchronized results in exploration of the search space while allowing localized searches.

4. Numerical Considerations and Proposed Algorithm

4.1. Numerical Aspects

It is important to note that, classical numerical ODE solving schemes fail for the stochastic differential equations (SDE) as in Eq. (6) because of the additional diffusion term due to the noise [10]. Many efficient numerical schemes has been introduced to solve SDEs. In the sequel, the basic Euler–Maruyama scheme, i.e.

$$
\mathbf{x}_{i+1} = \mathbf{x}_i + f(t, \mathbf{x}_i) R dt + \sqrt{dt} g(t, \mathbf{x}_i) \sum_{i=0}^{R-1} \omega_i, \quad \mathbf{x}_0 = \mathbf{x}(0)
$$

has been used for simulation of SDEs of the form $d\mathbf{x} = f(t, \mathbf{x}) dt + g(t, \mathbf{x}) \, d\mathbf{w}$, where $dt > 0$, $\lambda \triangleq x(kRdt)$, $R \in \mathbb{N}^+$ and $\omega_i$ are Gaussian increments with zero mean and unit variance (i.e. $\omega_i \sim N(0, 1)$ where $N(m, s)$ is the Gaussian density with mean $m$ and variance $s$). This scheme is chosen since it is easier to implement for high dimensional systems despite the fact that it is slow compared to more advanced schemes [11].

4.2. Stochastic CLM Algorithm

The overall algorithm that is proposed for global optimization is the following:

**Stochastic CLM Algorithm:**

1. **Initialization:** Determine $q$, $R$, $\Delta T$, $T_{\text{max}}$, $x_{0}^{(i)}$, $y_{0}^{(i)}$, $\mathcal{T}_0$, $\mathcal{J}_i$, $N$, $\gamma(0)$, $\eta(0)$, $\alpha$, $\gamma$, $\eta$, $\beta$, $\gamma$, $\eta$, $\epsilon$, $\mathcal{J}$, $i = 1, \ldots, q$ and $j = 1, \ldots, n$ and calculate $G^{(i)}_{\psi}(0)$ using Eq. (10).

2. **Optimization:** For $0 \leq k \leq T_{\text{max}}$

   (a) Simulate Eq. (6) for $\Delta T \leq t < (k + 1)\Delta T$ using the Euler–Maruyama scheme.

   (b) If $\|x_i^{(k)} - x_{i+1}^{(k)}\| \leq \epsilon$ then END.

   (c) Clip the solutions $x_i \leq x_i^{(k)} \leq \mathcal{J}_j$, to limit the search domain.

   (d) Adaptation : Calculate $\gamma(k+1)$ by solving (4), $\eta(k)$ using (5), and $G^{(i)}_{\psi}(k+1)$ using (10).

   (e) Repeat (a)–(d) until convergence, or $(k + 1)\Delta T = T_{\text{max}}$

Note that the output can be obtained from the cells which are not excited with noise for high accuracy.

5. Illustrative Examples

The following examples are chosen from [3] in order to illustrate the usefulness and improvement over deterministic CLMs.

5.1. Example 1 : Double potential well

Consider the function $U(x) = x^4 - 16x^2 + 5x + 100$. The actual global minimum is located at $x = -2.9035$, and the function has another local minimum at $x = 2.7468$. Simulations for the deterministic CLM are done by choosing $q = 3, \Delta T = 0.2, \gamma = 0.2, \gamma(0) = 1, i = 1, 2, 3, \mathcal{J} = 2, \alpha = 1, \eta = 1, R = 8, \lambda = 2 \times 10^{-3}, \epsilon = 10^{-3}$. The initial conditions are chosen randomly. When all the initial conditions are larger than $x = 0.1567$, the deterministic CLM network converges to the local minimum. The stochastic CLM is simulated by setting $T_0 = 500, \mu = 0.99$ and $\gamma = 1$. For a typical run, the states, the cost values during the evaluation and the noise gain $G^{(i)}_{\psi}$ are shown in Fig. 1(a), (b) and (c), respectively. It can be seen that the deterministic cells (i.e. $[2, 3]$) synchronize at the local minimum, the stochastic cell performs exploration since the noise amplitude is not zero. Even when the global minimum is reached by the deterministic cells, the stochastic cell continues to explore the space due to the high noise gain, until all cells agree on the solution. Then, the noise
amplitude goes to zero, effectively decoupling the noise from the system. For this illustrative example, thousand runs with the initial conditions distributed randomly in the interval $[-5, 5]$, has been performed for both deterministic CLM ($T_0 = 0$) and SCLM ($T_0 = 500, \mu = 0.99$). The deterministic CLM could achieve the global minimum the at 526 of the runs. The SCLMs converged to the global minimum in 640 of the runs, indicating a clear improvement. Note that only the state of the network at the final time is considered. The performance could be improved further easily by adding more deterministic and stochastic cells in the network.

5.2. Example 2

The cost function

$$U(x) = \frac{a}{2n} \sum_{i=1}^{n} x_i^2 + 8n - 4n \left( \prod_{i=1}^{n} \cos(\beta_1 t_i) + \prod_{i=1}^{n} \cos(\beta_2 t_i) \right)$$

(12)

where $x = (x_1, x_2, \ldots, x_n)^T$, $a = 0.01$, $\beta_1 = 0.2$ and $\beta_2 = 1$ has been optimized in the domain $[-20, 20]^n$ for $n = 50$. The minimum is known to be located at $U(0)=0$. Note that Eq. (12) has many local minima that are close to each other and have almost the same cost, hence the conventional gradient descent or line search based optimization methods fail, even with multiple starts. The simulation parameters are selected as $q = 25$, $\mathcal{J} = \{1, 6, 11, 16, 21\}$, $N = 50$, $\Delta T = 0.4$, $\gamma = 0.1$, $\gamma_0 = 1$, $\eta = 10^{-3}$, $\eta_0 = 1$, $\eta_I = 10^3$, $R = 8$, $dt = 10^{-3}$, $\epsilon = 10^{-4}$, $T_0 = 10^5$, $\mu = 0.99$. The SCLM algorithm converged to the global minimum in 433 epochs with in the given accuracy while deterministic CLM failed to achieve the global minimum with the same settings.

6. Conclusions

In this paper, we have presented stochastic CLMs for global optimization of smooth cost functions. A cooling schedule that depends on the average costate energy has been proposed. It allows global exploration of the search space, and automatic termination of the cooling when a solution candidate is agreed among the stochastic CLM processes for global optimization of smooth functions.


References

Dependence of Synchronous Activities in Coupled Inhibitory Neurons upon Excitability Classes

Shigeki Tsuji¹, Tetsushi Ueta², Hiroshi Kawakami³ and Kazuyuki Aihara¹

¹Aihara Complexity Modelling Project, ERATO, JST, 3-23-5-201 Uehara, Sibuya-ku, Tokyo, 151-0064, Japan, and Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
²Center for Advanced Information Technology, Tokushima University, 2-1 Minami-Josanjima, Tokushima, 770-8506 Japan
³ Tokushima University, 2-24, Shinkura, Tokushima, 770-8501 Japan
Email: shigeki@aihara.jst.go.jp, tetsushi@ait.tokushima-u.ac.jp, hukugakk@honbu.tokushima-u.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—Electrical coupling has been discovered extensively between same type interneurons in the brain. They are also coupled by inhibitory chemical synapses. However, firing property of single cell and the relationship between collective activity in interneuronal circuits and different functions of their coupling have not been revealed perfectly. To clarify above problems, we investigate bifurcation phenomena in coupled neuron models which show two excitabilities, Type I and Type II, and compare synchronous activities between two cases. As a result, we show in- and anti-phase solutions and complex behavior by the interaction of two different couplings in each system. We also show that coupled Type II neurons can easily synchronize than Type I but they are greatly affected by the inhibitory connection.

1. Introduction

Electrical coupling has been discovered extensively between same type interneurons in the neocortex, hippocampus, thalamus, and so on [1]. Moreover, they are often coupled by bi-directional or uni-directional inhibitory synapses, and exhibit rich phenomena by the effects of these couplings. The interplay of two different functions, voltage dependency and short-term plasticity, can develop advanced information processing in the cortical network. However, the relationship between the collective activity of interneurons and different functions of their coupling has not been revealed perfectly. In addition, many types of interneuron exist in the brain, and they show different activities. Detailed classification of such interneurons and the firing property of individual type, e.g., fast-spiking (FS) interneuron, have remained unclear.

The goal of our study is to clarify the influence on “system-level” dynamics resulting from the difference of firing property of single cell and the relationship between synchronous activities in interneurons and the effects of two different functional synapses. In this paper, we firstly adopt a single neuron model, two-dimensional Hindmarsh-Rose (2DHR) type model, with two types of spike generation mechanism, Type I and Type II. By using the bifurcation theory, we investigate synchronous activities in coupled neurons interconnected by electrical and inhibitory synapses, and thereafter identify parameter regions exhibiting various synchronous oscillations. We also compare the difference of synchrony between coupled Type I neurons and coupled Type II neurons. From these results, Type II neurons can easily synchronize than Type I neurons. The effect of inhibitory synapse is remarkably different between each type due to the difference of the excitability, especially, Type II neuron is greatly affected by the inhibitory connection, although electrical synapse enhances the in-phase oscillation in both types.

Finally, we examine spatio-temporal behavior in the large scale network which consists of homogeneous 2DHR type neuron models interconnected by electrical and inhibitory synapses. In the parameter region of two coupling coefficients, spatio-temporal complex behavior, irregular switching between synchronous and asynchronous oscillations, are observed extensively. They also coexist with synchronous firing in both networks. The firing frequency in the network of each type is obviously different, and hence each network might have a different functional role in local interneuronal networks.

Figure 1: Schematic diagram of coupled neurons connected by both electrical and bi-directional inhibitory synapses.
2. 2DHR-type neuron model

To investigate synchronization phenomena in coupled system interconnected by electrical and inhibitory synapses, we firstly consider a simple neuron model described as follows[2]:

\[
\begin{align*}
\frac{dx}{dt} &= c \left( x - \frac{x^3}{3} - y + z \right) \\
\frac{dy}{dt} &= \frac{x^2 + dx - by + a}{c}
\end{align*}
\]

(1)

where \(x\) and \(y\) denote the cell membrane potential and a recovery variable, respectively. \(z\) represents the external stimulus. This model is equivalent to the two-dimensional Hindmarsh-Rose model[3], if we consider the substitution of state variable, \(y \leftrightarrow -y\).

It is well known that 2DHR-type neuron model shows two kinds of excitability[4], Type I and Type II excitability associated with saddle-node and Hopf bifurcations, respectively. In fact, this model shows both excitability by setting certain parameter sets[2]. Hence, this model has the possibility to be able to represent the excitability of some neurons. However, which excitability does FS interneuron, common type interneuron coupled by both electrical and inhibitory chemical synapses, exhibit? This question has remained elusive. In recent theoretical studies[5, 6], they adopted the model which shows Type I excitability, but some experimental studies suggest that FS interneuron exhibits Type II excitability[7]. Therefore, in the following section, we adopt both excitabilities and compare synchronous phenomena in each case.

3. Bifurcations in inhibitory 2DHR-type neurons coupled by gap-junction

In this section, we investigate bifurcation phenomena in a fundamental system that two neurons are interconnected by electrical and bi-directional inhibitory synapses as shown in Fig.1.

The equations are

\[
\begin{align*}
\dot{x}_1 &= c_1 (x_1 - x_1^3/3 - y_1 + z_1) \\
&\quad + g_{gap}(x_2 - x_1) + g_{syn} s_1 (x_{syn} - x_1)) \\
\dot{y}_1 &= (x_1^2 + d_1 x_1 - b_1 y_1 + a_1)/c_1 \\
\dot{s}_1 &= \alpha (1 - s_1)/(1 + \exp(-x_1/0.1)) - \beta s_1 \\
\dot{x}_2 &= c_2 (x_2 - x_2^3/3 - y_2 + z_2) \\
&\quad + g_{gap}(x_1 - x_2) + g_{syn} s_2 (x_{syn} - x_2)) \\
\dot{y}_2 &= (x_2^2 + d_2 x_2 - b_2 y_2 + a_2)/c_2 \\
\dot{s}_2 &= \alpha (1 - s_2)/(1 + \exp(-x_2/0.1)) - \beta s_2
\end{align*}
\]

(2)

where \(g_{gap}\) and \(g_{syn}\) are the maximal electrical and synaptic conductances, respectively. For the inhibitory synaptic transmission, we adopt the first-order kinetics equation, \(s_{1,2}\), with a sigmoidal function. Although this model is not equipped with biophysically meaning, we can qualitatively investigate various synchronous phenomena and the effects generated by some important parameters in coupled system, e.g., the effects of the intensity of each coupling or the variation of firing frequency derived from the decay time constant of synaptic transmission.

Here, we suppose that both single neurons have the same internal parameters, i.e., they exhibit the same excitatory type. We use two parameter sets, \(a_1 = a_2 = 0.42, b_1 = b_2 = 1.0, c_1 = c_2 = 3.0, d_1 = d_2 = 1.8\) and \(a_1 = a_2 = 0.88, b_1 = b_2 = 1.0, c_1 = c_2 = 3.0\),
Coupled Type I neurons Coupled Type II neurons
-2
-1
 0
 1
 2
 0  100  200  300  400  500
x1, x2  →  t  →
-2
-1
...
enclosed by each set. The coupling coefficient of electrical
synapse does not influence the form of these curves. In
neurons, we compute the bifurcation diagrams in the
resting potential of each type, and the rise and decay time
lizes the in-phase solution via period-doubling bifurcation
in this diagram. However, the inhibitory coupling destabi-
junction, an in-phase solution (Fig.3-(a)) exists extensively
the same constant current
excitability, respectively[2]. Additionally, we assume that
inhibitory synapses. The region exhibiting the anti-phase
solution (Fig.3-(b)) is generated by both gap-junction and
the electrical coupling. On the other hand, the anti-phase
neurons) and
z
−
1
firing state with same frequency (≈ 35[Hz]) in both cases.
The synaptic reversal potential is set to
syn
−2
0
0
2 (for Type II neuron) less than the
0
5 (for coupled Type II neurons)
0
749
2
z
1
2
= 0.384 (for coupled Type II neurons)
0
and
β
= 0.05, respectively.

To investigate the interaction between two coupling co-
coefficients in coupled Type I neurons and coupled Type II
neurons, we compute the bifurcation diagrams in the
plane of each case as shown in Fig.2. In these bifurca-
tion diagrams, Gi, Ii and Pf i represent the tangent, period-
doubling and pitchfork bifurcations, respectively. i is a
nominal number. S in and S anti indicates the region where
the stable in-phase and anti-phase solutions exist, respec-
tively. In coupled Type I neurons, by the effect of gap-
junction, an in-phase solution (Fig.3-(a)) exists extensively
in this diagram. However, the inhibitory coupling destabi-
lizes the in-phase solution via period-doubling bifurcation
I2, but its effect is decreased by increasing the intensity of
the electrical coupling. On the other hand, the anti-phase
solution (Fig.3-(b)) is generated by both gap-junction and
inhibitory synapses. The region exhibiting the anti-phase
solution is shrank by increasing the parameter
. Moreover,
in the right-side region than
G4 curve, odd firing solu-
tions (Fig.3-(c)) generated by strong inhibitory coupling
exist, independent of the intensity of the electrical cou-
pling. Although two neurons are connected by fully and
symmetrically coupling and have the same internal param-
eters, the in-phase and anti-phase solutions are destabilized
by each pitchfork bifurcation
Pf i and
Pf 2, and then two sets
of solutions located at a symmetrical position occur simul-
taneously. Additionally, they change to chaotic phenomena
(e.g., Fig.3-(d)) via a period-doubling cascade, but such re-
gions are very small.

In coupled Type II neurons, in-phase solution (Fig.3-(a))
widely exists in the diagram than the case of Type I. The
anti-phase solution (Fig.3-(b)) exists in very small region
(weakly electrical and inhibitory coupling), and this solu-
tion is always coexistent with the in-phase solution. There-
fore, Type II neurons can easily synchronize than Type I
neurons. Odd firing solutions (Fig.3-(c)) exist in the area
enclosed by bifurcation curves
G4 and
I4, and greatly af-
flect the interaction of two coupling. The in-phase solu-
tion is destabilized via period-doubling bifurcation
I5, I6, and
I7. Moreover, this solution also destabilized via the
sets of bifurcation curves, G1 and I1, and G2 and I2, and
chaotic synchronization (Fig.3-(d)) exist in two regions
enclosed by each set. The coupling coefficient of electrical
synapse does not influence the form of these curves. In
other words, these bifurcations exist in the reduced system, self-inhibitory Type II neuron. Therefore, Type II neuron shows chaotic phenomena only by the self-inhibitory input, although Type I neuron does not.

Finally, we study the dynamical behavior in large scale network of same type neurons interconnected by electrical and inhibitory synapses. This system is two-dimensionally lined up with four neighbors coupling (50 × 50 neurons).

In this paper, however, we show relative spike timing and phase (φ) histograms in neighboring 10 pairs (Fig.4). In both of Type I network and Type II network, spatio-temporal complex behavior are widely observed in the parameter region of two coupling coefficients when the initial states of each neuron are chosen randomly (data not shown), but the 2DHR-type neurons connected by only weakly gap-junctional coupling show a relaxed in-phase synchrony. These network, however, can enhance synchrony by increasing the coupling coefficient of gap-junction. Moreover, Type II neurons obviously show a high degree of synchronous activity than Type I neurons. Inhibitory coupling decreases the degree of synchrony ϕ = 0.0 and enhances asynchrony toward ϕ = ±0.5, but sharpens synchrony than only weak gap-junctioonal coupling. Such a bistability is observed in coupled real interneurons, and this network may play an important role to generate and disrupt the cortical representation by irregular switching between synchronous and asynchronous (antisynchronous) firings.

4. Conclusions

We investigate bifurcation phenomena in coupled Type I neurons and coupled Type II neurons interconnected by inhibitory and the electrical coupling, and clarify that many firing patterns, including both the in-phase and anti-phase solutions, are generated by the interaction of each coupling. Coupled Type II neurons can synchronize than coupled Type I neurons, but show various firing patterns only by changing the inhibitory synapse, e.g., chaotic synchronization. In the large scale network of each type, the complex spatio-temporal dynamics, irregular switching the synchronous and asynchronous firing, are observed extensively in the parameter sets of two coupling coefficients. From these results, gap-junctional coupling can produce partially synchronous firing via the propagation of sub-threshold oscillations, while this coupling can also promote the effect of inhibition when inputs from neighbor cells have a certain time lag. Hence, this network may be able to switch between synchronous and asynchronous firing due to the input from other network, e.g., regular-spiking (RS) neurons. Moreover, interneurons may switch synchrony and asynchrony between these neurons in other network, effectively. These results may indicate the additional variety of information representation by the interaction of two different couplings in interneuronal networks. Type II neuron can synchronize than Type I neuron, but Type II neuron is greatly affected only by the change of inhibitory synapse. Therefore, heterogeneous Type II neuronal network may show more various firing activities including the synchronous firing.

Figure 4: Relative spike timing and phase (φ) histograms in 10 pairs. φ = 0.0 and φ = ±0.5 implies in-phase and anti-phase respectively. In each case, we fix the same parameter set of coupling coefficients, g_{gap} = 0.1 and g_{syn} = 0.1.

References


Synchronous Activities in Neuronal Networks Composed of Homogeneous and Heterogeneous Neurons

Kunichika Tsumoto†, Hiroyuki Kitajima‡ and Kazuyuki Aihara†

†Aihara Complexity Modelling Project, ERATO, JST; and Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan.
‡Dept. of Reliability-based Information Systems Engineering, Faculty of Engineering, Kagawa University, 2217-20, Hayashi-cho, Takamatsu 761-0396, Japan

Email: tsumoto@aihara.jst.go.jp, kitaji@eng.kagawa-u.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—We investigate bifurcations in coupled two Morris-Lecar neurons with the type I and II excitabilities. First, we analyze the homogeneous (coupled same types) neurons with the electrical synapse when one of the external DC current and the synapse conductance is varied. In this case the type II neurons have wider parameter regions of synchronous firing. For both types firing frequency is increased as the DC current is increased. Next, the heterogeneous neurons with the chemical synapses is studied and we show that the global bifurcation structure is decided by the type of the neuron with varied DC current.

1. Introduction

Neocortical interneurons, exclusively inhibitory neurons, exhibit heterogeneous morphology and electrophysiological properties[1]. By these differences they are categorized into several classes. The interneurons belonging to the same class form a inhibitory network with both electrical and chemical synapses[2]. The electrical synapses enable a pair of interneurons to fire in highly precise synchrony. The inhibitory networks composed of interneurons are now considered to be the controllers of several brain rhythms.

According to the difference of firing responses to steady current stimulation, neuron and dynamical models of spike generation are classified into two types, type I and type II[3, 4]. It has been shown by some experimental works[5, 6, 7] that regular-spiking pyramidal cells and fast-spiking nonpyramidal cells, a class of interneurons, exhibit properties of the type I and type II excitabilities, respectively. Thus, we should discuss how the differences in the membrane excitabilities of the interneurons and the pyramidal neurons or in the synaptic properties affect synchronous activities emerging from the networks of interneuron-interneuron or interneuron-pyramidal neuron pairs.

In the present study, we investigate synchronized oscillations observed in two Morris-Lecar (M-L) neurons[8] interconnected via chemical or electrical synapses. Since the property of the membrane excitability can be controlled by only one parameter in the M-L model[9], this model has an advantage in considering the effect of the difference between two types of the membrane properties on the synchronous activity. By numerical bifurcation analysis, we identify parameter regions in which several kinds of synchronous oscillations occur in the homogeneous neuron pairs. We also show that the decay time constant of the chemical synapses play an important role in the generation of synchronous activities in the heterogeneous pair of type I and II neurons.

2. Coupled Morris-Lecar Equations

In the presence paper, we study various synchronization phenomena observed in two-coupled Morris-Lecar (M-L) neurons. The dynamical system that we consider here (Fig. 1) is described as follows:

\[ \frac{dV^{[i]}_i}{dt} = \frac{-\bar{g}_{Ca}M^{[i]}_i(V^{[i]}_i - V_{Ca}) - \bar{g}_K N^{[i]}_i(V^{[i]}_i - V_K)}{C_M} - \bar{g}_L(V^{[i]}_i - V_L) + I^{[i]}_\text{ext} + g^{[i]}_\text{syn}, \quad i = 1, 2. \]

where \( V^{[i]}_i \), for \( i = 1, 2 \), is the membrane potential of the \( i \)th neuron, \( N^{[i]}_i \in [0, 1] \) is the activation variable for \( K^+ \)

Figure 1: Neuron model coupled via chemical (\( g_{\text{syn}} \)) electrical (\( e_{\text{syn}} \)) synapxes. We define the driving neuron as having varied external DC current. If the two neurons have the same (resp. different) type of the membrane excitability, i.e., the type I-I or the type II-II (resp. the type I-II or the type II-I) pair, we call homogeneous (resp. heterogeneous) neurons.
and $t$ denotes the time measured in milliseconds. $V_{Ca^+}$, $V_K$, and $V_L$ represent equilibrium potentials of $Ca^{2+}$, $K^+$, and leak currents, respectively, and $g_{Ca^+}$, $g_K$, and $g_L$ denote the maximum conductances of corresponding ionic currents. $I^{[i]}_{ext}$ is the externally applied DC current that is assumed to be temporally constant as a parameter.

In Eq.(1), $M^{[i]}_0$, $N^{[i]}_i$, and $\tau^{[i]}_N$, respectively, are defined as follows:

$$M^{[i]}_0 = 0.5 \left[ 1 + \tanh \left( \frac{(V^{[i]}_i - V_1)}{V_2} \right) \right],$$

$$N^{[i]}_i = 0.5 \left[ 1 + \tanh \left( \frac{(V^{[i]}_i - V_3)}{V_4} \right) \right],$$

$$\tau^{[i]}_N = \frac{1}{\phi \cosh \left( \frac{(V^{[i]}_i - V_3)}{2V_4} \right)},$$

see Refs.[8, 9] for more detailed description.

The synaptic current $I^{[i]}_{syn}$ for $i = 1, 2$, is a voltage-dependent function defined as follows:

$$I^{[i]}_{syn} = -g_{syn}(V^{[i]}_i - V_{syn})s^{[i]}(t) - e_{syn}(V^{[i]}_i - V^{[i]}),$$

where $V_{syn}$ indicates the synaptic reversal potential, and $g_{syn}$ and $e_{syn}$ are the maximum conductance of chemical and electrical synapses, respectively. The function $s(t)$ is defined by

$$s^{[i]}(t) = \frac{1}{\tau^{[i]}_1 - \tau^{[i]}_2} \left( e^{-\left( t - \tau^{[i]}_1 \right) / \tau^{[i]}_2} - e^{-\left( t - \tau^{[i]}_2 \right) / \tau^{[i]}_1} \right) \Theta(t),$$

where $\tau^{[i]}_1$ and $\tau^{[i]}_2$ are the rise and decay time of the synapse, respectively, and $\Theta(t)$ is the Heaviside function: $\Theta(t) = 0$ for $t \leq 0$ and $\Theta(t) = 1$ for $t > 0$. Note that the integral of $s(t)$ over time is normalized to 1. In Eq.(6), $\tau_d$ denotes a time delay for the synaptic transmission. Namely, the firing information of a neuron transforms to another neuron with the time delay $\tau_d$.

3. Results and Discussions

In the present paper, we investigate how differences of membrane property of neurons affect the synchronous activities in the M-L neuron pairs coupled via either chemical or electrical synapses.

We now treat $I^{[2]}_{ext}$ as a main control parameter, and analyze two-parameter bifurcations in the parameter plane of $I^{[2]}_{ext}$ and $g_{syn}$ or $e_{syn}$ for both type I and II neurons. These types can be switched by the value of the parameter $V^{[i]}_3$ in Eqs.(3) and (4)[9]. We fix $V^{[1]}_3$ as 12 [mV] and 2 [mV] for the type I and the type II neurons, respectively. In addition, to set a natural frequency of the driven neuron as 20 [Hz], the values of $I^{[1]}_{ext}$ for the type I and the type II neurons are fixed at 73.67 and 78.55 [µA/cm$^2$], respectively. The other parameter values in Eqs.(1)-(4) are listed in Table 1[9].

3.1. Electrical Synapse ($g_{syn} = 0$)

By changing the value of one external DC current denoted by $I^{[2]}_{ext}$, we can observe synchronous firing oscillations with various kinds of frequency-locking patterns.

For representing the frequency-locking patterns, we introduce the notation: “1 : $n$ mode-locking” is for frequency locked oscillation with $n$ spikes for the driven neuron. For example, the 1 : 2 mode-locking corresponds to the synchronous firing oscillation that the driving neuron produces two spikes during the generation of one spike in the driven neuron.

Bifurcation diagrams and average frequencies for firing oscillations observed in each pair of the type I and the type II neuron are shown in Fig.2. In the bifurcation diagrams, the regions filled by blue and red colors denote the parameter regions in which stable 1:1 and 1:2 mode-locking oscillations exist, respectively. In the gray regions we can observe a 1:0 mode-locking oscillation indicating a firing state only for the driven neuron.

From these figures we can see the following things: (1) Arnold-tongue-like structure surrounded by the saddle-node bifurcations can be observed; (2) In the region filled by blue in the bifurcation diagrams, the average frequency of the neuron pair is increased as the DC current $I^{[2]}_{ext}$ is increased; (3) For the type I neurons, a firing response with high frequency(∼ 26 [Hz]) is observed for small values of $I^{[2]}_{ext}$ and large values of $e_{syn}$ (see the right panel in Fig. 2(a)). This is the opposite property to an input-output relationship consisting of $I^{[2]}_{ext}$ and the firing frequency in the single neuron with the type I excitability [9]; and (4) The type I neuron pair have the wider firing frequency band(6 ∼ 26 [Hz]), while the type II pair have wider parameter regions in which the stable synchronous firing oscillations exist.

3.2. Chemical Synapse ($e_{syn} = 0$)

3.2.1. $\tau_1 = \tau_2 = \tau$

When $\tau_1 = \tau_2 = \tau$, we use the $\alpha$-function ($\alpha(t) = (t/\tau)e^{-\left( t/\tau \right)}\Theta(t)$) on behalf of the function $s(t)$ defined by Eq.(6). We show bifurcation diagrams obtained by the brute-force method for homogeneous and heterogeneous
neuron pairs coupled by only chemical synapses with $\tau = 1.0$. In this case also Arnold-tongue structure surrounded by the saddle-node bifurcations is observed. Comparing Fig.3(a) with Fig.3(d) (homogeneous case), the coupled type I neurons have more kinds of synchronized states (more colors in black regions), this is coincident with the case of the M-L neurons with a periodic input[10]. In the regions of large values of $I_{\text{ext}}^{(2)}$ and small values of $g_{\text{syn}}$ colored by light blue, when the driving neuron is type I (Figs.3(a) and (c)), only the driven neuron produces spikes. In coupled heterogeneous neurons if the driving neuron is the same type, then the global bifurcation structure is kept (Figs.3(a) and (c), and (b) and (d)). Detailed numerical bifurcation analysis is one of our open problems.

3.2.2. $\tau_1 \neq \tau_2$

In order to study the effect of the raise and the delay time constant of synapses on synchronous firing, we show a bifurcation diagram in Fig.4 for Eqs.(1) with $\tau_1 = 0.5$ [ms] and $\tau_2 = 7.0$ [ms]. Comparing Fig.4 and Fig.3(c) the long decay time constant suppress the several periodic synchronized firing in the regions of non-periodic states. The region of the fundamental synchronized states is also narrow for Fig.4.

4. Conclusion

We investigate bifurcations in coupled M-L neurons either electrical or chemical synapses. In electrical coupling the type I neurons have wider firing frequency than that of the type II neurons. Moreover, the parameter regions of existence of synchronized solutions are narrow for the type I neurons. In chemical coupling we compare the bifurcation structure for homogeneous and heterogeneous neurons. Regardless of homogeneous or heterogeneous global bifurcation structure is decided by the type of the driving neurons. In both cases we observe Arnold-tongue structure formed by the saddle-node bifurcations.

The study of both electrical and chemical coupling neurons and detailed bifurcation analysis changing the several

Figure 2: (Left) Bifurcation diagrams and (Right) average frequency of the network consisting of the neuron pair for electrical synapse coupling. In the left panels, blue, red, and gray regions indicate the existence of the 1:1, 1:2, and 1:0 mode-locking oscillations, respectively. Thick solid and thin dashed curves represent the saddle-node bifurcation and the period-doubling bifurcation, respectively. In the right panels, the average frequency for non-periodic states is approximated by obtaining an average period for 100 interspike intervals of the driven neuron.
Figure 3: Bifurcation diagrams for homogeneous and heterogeneous neurons. The regions are colored by its period of synchronous firing oscillations observed in each region. In black regions we observe non-periodic attractors. We change the value of the DC current of the type I neurons ((a) and (c)) and the type II neurons ((b) and (d)).

Figure 4: Bifurcation diagram for heterogeneous Type II - Type I neuron pair described by Eq. (1) with $\tau_1 = 0.5$ [ms] and $\tau_2 = 7.0$ [ms]. The meaning of colors is the same as Fig. 3.

Synaptic time constants are our open problems.

References


A Study on a Neural Network Model of Theta Rhythm and Sequential Memory
Youhei Nakamura† and Kazuyuki Aihara‡

† Graduate school of Information Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
‡AIHARA COMPLEXITY MODELLING PROJECT, ERATO,
Email: youhei@sat.t.u-tokyo.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract— A group of neurons generate rhythmic activity. The theta rhythm is one of this kind of activity, which consists of sinusoidal 4-12 Hz waves. This can be observed in the hippocampus when an animal explores a certain environment or when it sleeps with rapid eye movement. It is considered that the theta rhythm plays an important role in selecting and strengthening synaptic connections.

We show that the theta rhythm in a group of neurons can be constructed from a subthreshold oscillation of each neuron. We construct a neural network model of theta rhythm and shows that the dynamics of neural network changes from irregular spiking to regular spiking due to a slight change of input current in a vicinity of threshold potential for the spiking.

Sequential memory can be stored using the transition from an irregular spiking state to a regular spiking state. We show that our neural network model of theta rhythm can store sequential memory.

1. Introduction
In the brain, a group of neurons sometimes co-operate with each other and shows a rhythmic activity as a whole. Theta rhythm is one of this kind of activities, which consists of sinusoidal 4-12 Hz waves [1],[2]. The theta rhythm can be observed in the hippocampus by using EEG when an animal explores a certain environment or when it sleeps with rapid eyes moving (REM sleep). It is considered that the theta rhythm plays an important role in selecting and strengthening the information. A place cell is a neuron that codes a spatial location. When a rat moves around an environment, the firing rate of place cells depend on the location of the rat. Thus place cells form a place fields.

It has been proposed that the spiking time of the place cells relative to the phase of the theta rhythm provides the additional information about the location the rat exists. Each time the rat enters the place field, the place cell begins to fire at the same phase, and over the next five to ten cycles of the theta rhythm the phase of firing gradually gets faster up to 360 degrees. This phenomenon is called phase precession. While the rat is running along the truck, the phase of the place cell firing is more correlated with the rat’s location within the place field than the time that has passed since it entered the place field. The phase precession should be considered as a kind of temporal compression of the timescale of behavioral sequence into the synaptic timescale.

A number of neural mechanisms are proposed for theta rhythm and sequential memory, based on the phase precession [3],[4]. However, the mechanisms of phase precession have not been known. We proposed that theta rhythm changes the efficacy of synaptic modulation by using the change of spiking regularity.

2. Model and Method

2.1 Classification of neuron and bifurcation
The mechanisms of a spike generation were studied in detail and modeled by Hodgkin and Huxley. They have also found that neurons can be classified by its spiking property; Class 1 and Class 2 [5]. Class 1 neuron can generate spikes with arbitrarily low frequency depending on the strength of the applied current. Class 2 neuron generates spikes in a certain frequency band that is relatively insensitive to changes in the strength of the applied current. Classes of excitability occur because neurons have different bifurcations of resting and spiking states [6].

The bifurcations from a resting state to spiking state are mainly classified into four bifurcations: saddle-node bifurcation, saddle-node on invariant circle (SNIC) bifurcation, subcritical Andronov Hopf bifurcation, and supercritical Andronov-Hopf bifurcation.

The equilibrium of neurons is classified based on whether a neuron is an integrator or a resonator. The neuron is an integrator when the equilibrium that corresponds to a resting state is a node. The neuron is a resonator when the equilibrium is a focus. The integrator spikes with arbitrary large latency, whereas the resonator exhibits a damped oscillation.

2.2 Izhikevich model
We numerically implement our model of theta rhythm and by using the model proposed by Izhikevich (2003) [6][7]. From now on, we call this model Izhikevich model. The Izhikevich model can be written as

\[ C\dot{v} = k(v-v_r)(v-v_l) - u + I \]
\[ \dot{u} = a(b(v-v_r)-u) \]

if \( v \geq v_{peak} \), then \( v \leftarrow c, u \leftarrow u + d \).

where \( v \) is the membrane potential, \( u \) is the recovery variable which accounts for the activation of \( K^+ \) ionic channels and inactivation of \( Na^+ \) ionic channels, \( C \) is the membrane capacitance, \( v_r \) is the resting membrane potential.
potential, and $v_t$ is the instantaneous threshold potential. In the Izhikevich model, if $v$ skips over $v_{peak}$, then $v$ is reset to $c$ and $d$ is added to $u$. This resetting represents the cut of absolute refractory period.

In our model, we use two types of neurons: regular spiking (RS) neurons and fast spiking (FS) neurons. RS neurons are chosen in Class 2 regime. Injection of current above 40pA changes the resting state to a stable focus and results in damped oscillations of the membrane potential. The basin of attraction for the focus is bounded by the stable manifold to the saddle. The stable manifold makes a loop and becomes a homoclinic orbit when $I=124.5$ pA. This homoclinic orbit gives birth to a spiking limit cycle attractor, and the stable focus and newly appeared spiking limit cycle co-exist when $I=125$ pA. In this region, neurons exhibit stuttering behavior. As the input current $I$ increases, the sum of eigenvalues becomes positive. When the stable manifold makes another homoclinic orbit, it gives birth to an unstable limit cycle. This unstable limit cycle shrinks toward the stable focus, and results in a subcritical Andronov-Hopf bifurcation.

The relationships between input current and frequency of stable focus, unstable limit cycle, and spiking limit cycle are shown in Fig.1.

![Figure 1. Frequencies of the equilibrium (blue) and periodic spiking (red) depending on input current. The pink dot line is a frequency of unstable limit cycle. The green line indicates the point where stable focus and unstable limit cycle coalesce and annihilate.](image1)

2.3 Network model of theta rhythm

We hypothesized that a theta rhythm is formed by the property of stable focus. Therefore, we used the Class2 parameters of RS neurons, whose equilibrium bifurcates by the subcritical-Hopf bifurcation. In this RS model, the points that are attracted by an equilibrium converge to the equilibrium with rotating around the equilibrium.

Therefore, the membrane potential decays to the resting potential with a damped oscillation, which is called depolarized after potential (DAP) after a spike (Fig 2).

The next spike tends to occur at the peak of DAP because the membrane potential is large than. A theta rhythm is formed by DAP and subthreshold oscillation near the equilibrium.

Based on this assumption, we construct a network model of theta rhythm, which consists of 80 RS neurons and 20 FS neurons. The number of synaptic connections from RS to RS, RS to FS, FS to RS, FS to FS are $16 \times 80$, $4 \times 20$, $4 \times 80$, $4 \times 20$ respectively, and are randomly arranged.

Synapses which happened to be connected with themselves were removed.

![Figure 2 Depolarized after potential (left: time course of V, right: phase plane)](image2)

For quantifying the spiking irregularity and a degree of synchrony in the network, we introduce the coefficient of variation (CV) of all inter-spike intervals and the peri-stimulus time histogram (PSTH). The coefficient of variation CV is defined as

$$C_V = \frac{\sqrt{\frac{1}{n} \sum_{i=1}^{n} (T_i - \overline{T})^2}}{\overline{T}}.$$  

(2)

where $T_i$ is the duration of the each interspike interval (ISI), $n$ is the number of ISIs for all the neurons, and

$\overline{T} = \frac{1}{n} \sum_{i=1}^{n} T_i$ is the mean ISI. The value of CV takes 1 for purely Poisson process, and 0 for a perfectly regular sequence. Thus CV indicates the global spiking irregularity for all the neurons in the network.

PSTH is the histogram of spikes for all the neurons in the network. If PSTH does not have sharp peaks, spikes between neurons in the network are weakly synchronized. If PSTH has a rugged shape, they are strongly synchronized. We use CV and PSTH for the estimating whether the firing patterns of neurons are generated by the subthreshold or by the periodic spiking.

2.4 Sequential memory

For studying the relationship between theta rhythm and sequential memory, we construct a network model with STDP. The sequential memory we assume here is the navigation system, when a rat moves on a straight road and remembers the path. When the rat moves straight from the right to the left, place cells that encode the place the rat stays receive an external input and fire. Each place cell has a place field that is aligned with a fixed interval.

Differences between timings when each neuron begins to receive the external currents are related to the distances between the place fields and the speed the rat moves. When the rat moves at constant velocity, two neurons whose place fields are different from each other begin to fire at different timing. This time difference depends on the speed and distances of place fields.

The basic concept of sequential learning is that place cells begin to fire with regular interspike intervals when the rat
is in place fields whereas they fire even if the rat is outside the place fields depending upon the theta rhythm that is made from the subthreshold oscillation of each neuron. When the input current is weak the rat is out of the place fields, the neurons fire sparsely. Therefore they can eliminate futile learning because intervals of spikes among neurons are broad and irregularly changes in a few cycles. If the input current is increased the rat is in the place fields, neurons fire regularly. In this state, interspike intervals are uniform and they can learn effectively.

The different onset time of regular spiking state results in the different phase firing in the theta rhythm. Sequential memory is achieved by this transition. Initial phase differences are given by the delay of input. We use four intervals, 1ms (in one cycle), 2.5ms, 5ms, and 10ms (over one cycle). In the case that total input delay is over one cycle, the spike sequence is divided by the theta rhythm, and the groups of neurons whose spike sequence is in a cycle are formed. We examine the changes of synaptic connections along with different onset interval. Synaptic connections from RS to RS, RS to FS, FS to RS, FS to FS we used are 48×80, 4×20, 4 × 80, 4× 20 respectively. These synaptic connections were randomly arranged.

3. Results

3.1 Network model of theta rhythm
An example of theta rhythm in our model is shown in Fig.3. The theta rhythm, we explained in the previous chapter, consists of sinusoidal 4-12 Hz waves whose amplitude is 1-2mV. In this sense, when I=126pA, a theta rhythm is formed. When I=135pA, neurons are nearly synchronized and spike irregularity is vanished. Therefore, we call this state a synchronized periodic spiking state.

The relationship between input current I and CV, and that between I and PSTH are shown in Fig.4. PSTH is gradually sharpened and more rugged as the input current increases. This means that the neurons in the network are gradually synchronized. In addition to the synchronization, the coefficient of variation CV for the interspike intervals for all the neurons decreases. This suggests that the neurons in the network spikes more regularly if the input current gets stronger.

Figure 3. Theta rhythm, I=126pA (up) and synchronized periodic spiking state, I=135pA (bottom).

![Figure 3. Theta rhythm, I=126pA (up) and synchronized periodic spiking state, I=135pA (bottom).](image)

Figure 4. PSTH and CV with different inputs

3.2 Sequential memory
The irregularity and synchrony of the firing pattern can change with the amplitude of input current. We used the transition from an irregular spiking state (I=126pA) to a regular spiking state (I=130pA), and examined whether sequential memory can be stored by varying an onset interval D of external input. The number of learning stages applied just once. During the learning stage, 4pA external inputs were inserted during 2ms. The results are shown in fig.5. When D=1ms, neurons were almost synchronized after the learning stage. The firing patterns of the network were fixed ever after the learning stage. When D=2.5ms, which corresponds to a cycle of theta rhythm, a sequential firing pattern were stored finely. When D=5ms and D=10ms, firing patterns got separated.

5.Conclusion
We have shown that a theta rhythm in an entire neural network is constructed from a property of subthreshold oscillation which is inherent in a single neuron. The neurons that exhibit a subthreshold oscillation are called resonators. We have also shown the important role of resonators, that has been inclined to be ignored by present researchers.
In the case of $D=1\text{ms}$

![Synaptic weight matrix](image1)

Synaptic weight matrix raster plot before/after leaning

$D=2.5\text{ms}$

![Synaptic weight matrix](image2)

Synaptic weight matrix raster plot before/after leaning

$D=5\text{ms}$

![Synaptic weight matrix](image3)

Synaptic weight matrix raster plot before/after leaning

$D=10\text{ms}$

![Synaptic weight matrix](image4)

Synaptic weight matrix raster plot before/after leaning

Figure 5. synaptic weight matrix (left) and firing pattern before/after learning

We have introduced the coefficient of variation $Cv$ as an indicator of irregularity in order to examine the relationship between the input current and irregularity of interspike intervals in the entire neural network. The value of $Cv$ gradually sinks down as the input current increases. Using $Cv$, we have found that internal dynamics of neural network shifts from “irregular, resting state-origin” dynamics to “regular, periodic spiking state-origin” dynamics due to the slight change of input current in a vicinity of threshold potential of neuron that is called peri-threshold. A spontaneous activity of neural network tends to become a peri-threshold state. Interspike intervals are irregular while a neural network stays in this state. Therefore, we have claimed that the irregularity of interspike intervals has a role in preventing synaptic connections in the network from getting strong or weak excessively due to the Spike-Timing-Dependent-Plasticity.

Finally, we have shown that sequential memory can be realized using this transition from an irregular spiking state to a regular spiking state. Plastic learning hardly occurs in an irregular spiking state because a difference of spike timings between neurons in the network changes irregularly. However, the delays of the spike timing between an arbitrary pair of neurons are constant in a regular spiking state. We assumed that sequential memory is stored when neurons receive an external input and change the state from an irregular to a regular state. We have shown that sequential memory is realized.

Acknowledgments

This research is partially supported by Grant-in-Aid for Scientific Research on Priority Areas -System study on higher-order brain functions- from the Ministry of Education, Culture, Sports, Science and Technology of Japan (17022012).

References


Entrainment of Circadian Oscillations under 24h Light-Dark Cycles

Gen Kurosawa†‡, Kunichika Tsumoto†‡, and Kazuyuki Aihara†‡

† Aihara Complexity Modelling Project, ERATO, JST
Komaba Open Laboratory, The University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan
‡ Institute of Industrial Science, The University of Tokyo
4-6-1, Komaba, Meguro-ku, Tokyo 153-8505, Japan
Email: kurosawa@aihara.jst.go.jp, tsumoto@aihara.jst.go.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—Organisms evolved to have circadian clocks against the abrupt change of environment. We and other authors previously demonstrated that large amplitude of light-dark cycles may induce the complex oscillations in the form of period-2 or chaos in the abundance of clock proteins. Using molecular models for circadian rhythms in Neurospora, we determined the effect of all-or-none change of light conditions in the morning phase, and also that in the evening phase on the entrainment of clock oscillations. Finally we discuss the physiological relevance of these results.

1. Introduction

Organisms show autonomous oscillations in the ambient change of environment. Circadian clock persists under constant dark conditions. Molecular clock systems underlie behavioral rhythms and metabolic rhythms. In Drosophila, from period gene (abbreviated as per gene), per mRNA is synthesized. Then, per mRNA is translated into PER protein which suppresses the transcription of per gene by indirectly inactivating the binding of transcription factor CLOCK (abbreviated as CLK) to the promoter region. This negative feedback regulation of clock genes is suggested to generate the sustained circadian oscillations for Drosophila and also for other organisms such as Neurospora, plants, and mammals [1-4]. Moreover sustained oscillations at cellular level cause the circadian oscillations at behavioral level.

A conspicuous property of the circadian clock is the entrainment of circadian oscillations by light-dark cycles (abbreviated as LD cycles). Period of free-running oscillations is different from 24h in many organisms. The free-running period of Drosophila is 24.3h, that of Neurospora is 21.5h, that of Phaseolus is 28h, and that of humans is 24.18h [5-8]. Thus, adjustment of circadian oscillations by using the light input is required for organisms to synchronize with the period of earth rotating rhythm. In Drosophila, PER protein forms complex with TIMELESS (abbreviated as TIM) protein in the cytoplasm. Heterodimer of PER and TIM enters the nucleus and then suppresses the expression of per gene and tim gene. Light destabilizes TIM protein in Drosophila. Subsequently cytosolic PER is decomposed because PER is unable to form stable complexes with TIM [9].

Since a seminal paper by Goodwin [10], many mathematical models have been proposed for the circadian oscillations in Drosophila[11-15], that in plants [16], that in Neurospora[17,18], that in cyanobacteria [19], and that in mammals [20,21]. Goldbeter [11] proposed the model for circadian rhythms in Drosophila in which PER protein is phosphorylated and then suppresses the transcription of per gene. These demonstrated sustained oscillations for certain parameter condition and the entrainment of circadian oscillations by 24h light-dark cycles. Gonze and Goldbeter [22] showed that complex oscillations in the form of period-2 and chaos under light dark cycles can occur in circadian rhythms model for Neurospora. In this article, we examine the effect of all-or-none change of period-forcing in the onset of light phase, and that in the offset of light phase, by using the same model of Neurospora. Finally, we discuss the physiological relevance of these results.

2. Neurospora Model

We first consider a minimal model for the gene-protein network generating the circadian rhythms in Neurospora [17,22]. In the model, we neglect the phosphorylation and dephosphorylation of protein. Moreover other clock genes are not incorporated which have been also identified as clock genes. The kinetic equation is:

\[
\begin{align*}
\frac{dM}{dt} &= \frac{v_{F_h}^{n}}{K_n + F_N^{n}} - \frac{v_{m}M}{K_m + M} \quad (1a) \\
\frac{dF_C}{dt} &= k_x M - \frac{v_{d}F_C}{K_d + F_C} - k_1 F_C + k_2 F_N \quad (1b) \\
\frac{dF_N}{dt} &= k_4 F_C - k_2 F_N \quad (1c)
\end{align*}
\]

where \( f_q \) mRNA (\( M \)), \( FRQ \) protein in cytosol (\( F_C \)), and \( FRQ \) protein in nucleus (\( F_N \)). First term of the right hand side of Eq. [1a] indicates the expression of \( f_q \) gene which decreases as \( FRQ \) protein accumulates in the nucleus. As schematized in Fig.1, \( f_q \) mRNA (\( M \)) is translated into \( FRQ \) protein (\( F_C \)) in the cytosol. \( FRQ \) enters the nucleus where \( FRQ \) suppresses the expression of \( f_q \) gene. Sustained oscillations occur for certain parameter condition. Oscillation with the period of 21.5h which is...
actually the free-running period of *Neurospora* occur for the parameter set proposed by Leloup et al. [16] (see caption for Fig.1).

Circadian oscillations of FRQ with the free-running period of 21.5h can be entrained by 24h light-dark cycles as $v_{rmax}$ increases (Fig.1). However, if we further increase $v_{rmax}$, complex oscillations occur and FRQ oscillations are not entrained by LD cycles (Fig.2). For $v_{rmax} = 2$, for instance, period-2 (Fig.3) oscillations occur. For $v_{rmax} = 3$ (Fig.4) chaos occur, and then the period of oscillations and the phase of the maximum value of FRQ protein under 24h light-dark cycles fluctuate. Period doubling bifurcation point is when $v_{rmax} = 2.636$ (Fig.2).

2.1 Entrainment of circadian oscillations of FRQ by 12h:12h LD cycles

In *Neurospora*, light acts by enhancing the expression of *frq* gene [23]. Expression rate of *frq* gene $v_x$ is fixed at the constant value $v_{rmax}$ during the light phase. We wish to determine the effect of $v_{rmax}$ on the entrainment of FRQ oscillations by 12h:12h light-dark cycles.

Heintzen et al. [24] demonstrated that accumulated *frq* mRNA is quickly degraded after transcription of *frq* is activated by light. This finding suggests that degradation rate of *frq* mRNA may decrease during light phase, or enhanced transcription of *frq* may be suppressed after the activation of transcription by light which are not incorporated in Eqs. [1] for simplicity.

If $v_{rmax}$ is smaller than 1.92, quasi-periodic oscillations occur and circadian oscillations of FRQ protein can not be entrained by 12h:12h LD cycles when the free-running period is 21.5h.
hand, the domain of entrainment by LD cycles extraordinarily enlarges for the latter case. To our surprise, when $T_r=4h$, FRQ oscillations are entrained if $v_{r_{max}}$ is in the range between 2 and 40. Moreover, the domain of entrainment by LD cycles enlarges as $T_r$ decreases. Thus complex oscillations in the form of period-2 or chaos are more likely occur if sustained period of the increase of transcription rate $v_{r_{max}}$ after the onset of light phase is shorter.

6. Discussion

An intriguing property of circadian rhythms is the circadian oscillations can be entrained by light-dark cycles. In Neurospora, light activates the transcription of frq mRNA and adjusts the period of oscillations. Activation of transcription rate of clock gene underlie the entrainment of circadian oscillations by LD cycles also in mammals. On the contrary, light activates the degradation of clock protein in Drosophila. Mathematical models demonstrated that complex oscillations of protein abundance in the form of period-2 or chaos can occur if the amplitude of period forcing is sufficiently strong [22,25,26].

In the present paper, we determined the effect of all-or-none change in the onset of light phase, and that in the offset of light phase. Three variable model for Neurospora displays entrainment by 12:12 LD cycles in a certain range of maximum value of transcription rate during light phase. As the entrainment strength increases, domain of entrainment appear, and then disappear again.

In Neurospora, accumulated frq mRNA is quickly degraded after transcription of frq is activated by light [24]. This indicates that degradation rate of frq mRNA may decrease during light phase, or enhanced transcription of frq may be suppressed after transcription is activated by light. In mammals, light also activates the expression of clock gene (i.e. mPer1) and then degraded during first several hours [27]. When amplitude of period-forcing is fixed at the constant value for first several hours and decreases for the rest of hours during light phase, domain of entrainment greatly enlarges in the model given by Eqs. [1]. An advantage of decreasing the transcription after light enhances the transcription rate may be to stabilize the entrainment over large range of light intensity.

References

[22] D. Gonze, A. Goldbeter, "Entrainment versus chaos in a model for circadian rhythms in *Neurospora* and *Drosophila*," 
rapid increase in frequency transcript," *Cell* vol.81, pp.1003-12, 1995.
*J. theor. Biol.* in press.
mammalian circadian clock is associated with rapid induction of mPer1 transcript," *Cell* vol.91, pp.1043-53, 
1997.
Dynamic Transition among Memories on Neurocomputer Composed of Amoeboid Cell with Optical Feedback

Masashi Aono and Masahiko Hara

Frontier Research System, RIKEN (The Institute of Physical and Chemical Research), Wako 351-0198, Japan.
E-mail: masashi.aono@riken.jp

Abstract—True slime mold Physarum polycephalum, the unicellular amoeboid organism having an oscillating cellular membrane, can be regarded as a biological oscillatory medium capable of spatiotemporal behavior-based information processing making reasonable decisions. To explore the organism’s decision-making in a dynamic environment, we let the organism act under optical feedback incorporating Hopfield’s neural network algorithm for associative memory. Our system can associate an input with one of suitable memories, as the organism having photoavoidance can search for the stable memory configuration by changing its planar shape inside a branched structure, so as to maximize the body area without being irradiated. Interestingly, even though no external perturbation is applied, the organism spontaneously starts to destabilize the once-recalled memory to seek another memory, as some branches of the organism grow under irradiation contrary to photoavoidance. By repeatedly switching between the stabilizing and destabilizing modes with spontaneous transition of the organism’s oscillation patterns, our system achieves dynamic transition among memories. The result suggests the unicellular organism’s potential to solve optimization problems on a phase space of the spatiotemporal oscillation patterns.

1. Introduction

The unicellular amoeboid organism Physarum polycephalum (Fig. 1A) is a multinucleated giant cell whose simple and homogeneous body structure consists mainly of a single extracellular gel layer (cellular membrane) and protoplasmic sol flowing through the intracellular channel network. Despite the absence of highly differentiated structure such as a central system, the organism exhibits sophisticated computational capacities, as Nakagaki and coworkers showed that the organism could solve a maze by connecting the minimum-length route between two food sources to achieve the most efficient nutrient absorption [1].

The organism’s dynamic behavior is mainly determined by the rhythmic contraction-relaxation cycle of the gel layer with a period of 1 to 2 min, derived from the collectively entrained oscillation of actin fibers. The shape deformation and locomotion are driven by the contraction-relaxation cycle as it induces streaming (shuttling) of protoplasm with periodic oscillation of body thickness. Furthermore, the periodic thickness oscillation involving phase lags generates phase waves, and information (strength and direction) on various stimuli, such as chemicals, temperature and light, received at a local site can be transmitted throughout the body by propagating a phase wave. In this sense, the organism is a fascinating research subject of spatiotemporal behavior-based information processing of biological oscillatory media.

To focus on the formation of the spatiotemporal oscillation patterns in well-formed boundary conditions, Takamatsu formed “living coupled oscillator systems” in which an individual Physarum is placed inside a microfabricated structure consists of several connected wells as oscillators [2]. Depending on the number and geometry of the oscillator connection (i.e., the width and length of channels connecting the wells to regulate the coupling strength), different multistable spatiotemporal patterns of thickness oscillation were observed, and spontaneous transition among the multistable oscillation patterns was reported. In this paper, we explore the functionality of the transition of the oscillation patterns for the organism’s survival, and report an illustrative embodiment utilizing the transition behavior for providing a dynamic feature in constructing an associative memory device. Our system is the first experimental implementation of a neurocomputer executed by a living amoeboid cell with optical feedback [3].

2. Methods

We fabricated a circuit structure (barrier) consisting of eight radially branched paths (Fig. 1A), to restrict the area of the organism’s activities on agar gel. We call the $i$ th path of the circuit “neuron $i$” ($i \in I = \{1, 2, \ldots, N = 8\}$). The organism’s shape is monitored using a video camera at each time interval of $\Delta t = 6$ sec, to evaluate each neuron’s state by digital image analysis. For each neuron $i$ at time $t$, whenever more than a quarter of the area of the $i$ th neuron is occupied by the organism’s branch, state 1 (active) is assigned as $s_i(t) = 1$, otherwise $s_i(t) = -1$ (inactive).

The branch of the organism degenerates (shrinks) when light stimulation is applied locally, as protoplasm is extruded from the irradiated site with light-induced contraction enhancement of actin fibers in the gel layer. Due to this photoavoidance, we can induce a neuron to take state $-1$ by irradiating the corresponding neuron. A projector
Figure 1: (A) Au-coated circuit made of plastic is placed on a 1% agar plate without nutrients. An aliquot (0.75±0.05 mg) of the amoeboid organism can act only inside the circuit where agar is exposed, because the organism averts the dried metal surface (scale bar=2mm). (B) Lateral inhibition rule for optical feedback: active state \( s_i(t) = 1 \) (indicated by hooked symbol) triggers light irradiation \( l_{i-1}(t + \Delta t) = l_{i+1}(t + \Delta t) = 1 \) (white light projected to yellow rectangle regions), where the boundary condition is periodic (scale bar=2mm). (C) All memory configurations consisting of rotation symmetric equivalents of the X- and Y-shaped configurations determined by the lateral inhibition rule.

connected to PC enables automatic irradiation control with projection of a monochrome image in which irradiated and nonirradiated regions are color-coded respectively white and black. We write \( l_i(t) = 1 \) if the light for the neuron \( i \) is turned on. Conversely, if the neuron \( i \) is allowed to be \( s_i(t) = 1 \), the light is turned off as \( l_i(t) = -1 \). In such cases, the neuron \( i \) is expected to be fully occupied because the organism grows concentrically or radially in principle (i.e., expands all of its branches). The irradiation pattern \( < l_1, l_2, \cdots, l_N > \), therefore, can lead the system configuration \( < s_1, s_2, \cdots, s_N > \) toward the ideal configuration \( < s'_1, s'_2, \cdots, s'_N > \) such that all neurons fulfill the “counteractive rule”. \( s'_i + l_i = 0 \).

The optical feedback implements a discrete form of Hopfield’s neural network algorithm [4] for associative memory in assuming the counteractive rule as follows:

\[
-l_i(t + \Delta t) = f(\sum_{j=1}^{N} w_{ij} s_j(t)),
\]

where \( w_{ij} \) is the weight assigned to the link from the neuron \( j \) to \( i \), and the step function \( f(\sigma) = 1 \) if \( \sigma > 0 \), otherwise -1. That is, if the right hand side of Eq. (1) requires the neuron \( i \) to be inactive \( (s_i(t + \Delta t) = -1) \) then the light \( i \) is turned on \( (l_i(t + \Delta t) = 1) \) and vice versa. Accordingly, the dynamics Eq. (1) establishes a certain rule to regulate how to update the irradiation pattern. With the change in the system configuration, the irradiation pattern is updated according to the rule at each interval, and again it induces further shape change of the organism. This iterative interaction between the system configuration and the irradiation pattern is the essence of our computing process [5].

In the above formalization, a memory is represented by a steady state (i.e., a stably fixed configuration). When the time evolution started from an input configuration reaches the steady state, the memory is regarded as recalled. If the organism takes the configuration such that the counteractive rule is satisfied for all neurons and the irradiation-updating rule gives no change in the irradiation pattern, the organism is no longer forced to reshape by irradiation and can fully expand its branches inside all nonirradiated neurons. Such a condition must be a comfortable environment for the organism. Thus, the system can stably maintain the configuration as a memory, unless the organism spontaneously breaks the counteractive rule.

If we represent the \( k \) th memory configuration as \( < m^k_1, m^k_2, \cdots, m^k_N > (m^k_i \in \{1, -1\} \text{ and } k \in \{1, 2, \cdots, M\}) \) and define the irradiation-updating rule by assigning all weights according to the Hebbian rule: \( w_{ij} = \sum_j m^k_i m^k_j \) if \( i \neq j \), otherwise 0, the dynamics Eq. (1) can be defined so that the network stores a number of arbitrary memories and associates an input with a suitable memory that is closest to the input. Adopting the Hebbian rule, however, makes the irradiation-updating rule complicated in analyzing experimental results. In the following experiment, we define the irradiation-updating rule in a simpler manner without spoiling the essence of the formalization.

We introduce the “lateral inhibition rule” (Fig. 1B) with the weight assignment \( w_{ij} = -1 \) if \( j \neq i = 1 \), otherwise 0. That is, if the neuron \( i \) is active \( (s_i(t) = 1) \), its adjacent neurons \( i-1 \) and \( i+1 \) are induced to be inactive by irradiation \( (l_{i-1}(t + \Delta t) = l_{i+1}(t + \Delta t) = 1) \). This irradiation-updating rule prohibits two adjacent neurons \( i \) and \( i+1 \) from being active simultaneously. We employed this rule because it presents highly symmetric 10 memories (Fig. 1C), that cannot be realized using the Hebbian rule.

3. Results

Figure 2 shows a typical computing process started from the input configuration \( < -1, -1, -1, -1, -1, -1, -1, 1, 1 > \) (Fig. 2A). As all branches followed the counteractive rule, the system recalled (reached) and stably maintained a configuration (Fig. 2B), which is one of memories closest to the input. Thus, our system successfully functioned as an associative memory device.

The stabilizing mode for the association, however, was spontaneously switched to the destabilizing mode even though no external perturbation was applied explicitly, as the newly-emerged branches 3 and 7 suddenly started to invade irradiated regions contrary to previous photoavoidance (Fig. 2C). While aggressive growth of the invading branch 7 was sustained under irradiation, the branch 8 was degenerated by irradiation according to the lateral inhibition rule, and the once-recalled memory (Fig. 2B) evolved into another memory (Fig. 2D). Spontaneous mode switch-
Figure 2: Time course of events in typical computing process. Red and blue pixels indicate that the thickness of the corresponding local site increased and decreased, respectively. (A) Input configuration. With the concentric growth, circular phase wave propagates from the center to the periphery. (B) First-recalled memory configuration (duration=4h). (C) Destabilization of memory B. Arrows indicate the growth directions of lumplike bulges invading irradiated regions. (D) Second-recalled memory (duration=1h). (E) Destabilization of memory D. (F) Third-recalled memory (duration=7h).

ing was repeated again (Fig. 2E). Consequently, the system achieved the dynamic memory transition by successively recalling three memories (Figs. 2B, D and F) within 16h. The organism’s volume was almost constant throughout the observation. In 20 experimental trials measured for an average of 14.5 h, an average of 3.25 memories (gross number including multiple recalls of an identical memory) were recalled per trial.

4. Analysis

Under equal irradiation conditions, in the memory destabilizing mode, some branches expand contrary to photoavoidance while others shrink as usual, whereas in the memory stabilizing mode, all branches keep away from irradiation. This spatiotemporal discrepancy of photoavoidance, more generally, flexible variability of stimulus-response, is essential for the survival of the organism required to search for food in a harsh environment, because it enables the organism surrounded by aversive stimuli to break the deadlock condition with enterprising responses. This feature allows us to consider that the organism can move into uncomfortable conditions in search of a better environment. We analyzed the organism’s oscillatory behavior to explore how the organism switches between the stabilizing and destabilizing modes.

As shown in Fig. 2, by binarizing the thickness oscillation of each local site into increasing (relaxing) and decreasing (contracting) phases, we can observe phase waves propagating vigorously throughout the body. As a phase wave enters and leaves a neuron (branch), the area of the domain where thickness is increasing (the number of red pixels) oscillates on a time scale of the contraction-relaxation period. In Fig. 3 we measured the fraction of this thickness-increasing area for each neuron i and the circuit’s center disc c, denoted respectively r_i(t) and r_c(t). To evaluate the substantive contribution of phase wave propagation to the growth/degeneracy (expanding/shrinking) movements of each branch, we calculated the phase difference d_i(t) ∈ [0°, 180°] between the two oscillating variables r_i(t) and r_c(t) (i.e., the net phase difference defined in [0°, 360°] is folded to half), because phase waves topologically propagate via the center disc. The growth/degeneracy movement was measured by the fraction of the area occupied by the branch i, denoted a_i(t).

Figure 3 shows that the newly-emerged branch 4 rapidly invaded the irradiated region by sustaining a large phase difference, which is close to the exact antiphase (i.e., a half period d_i(t) ≃ 180°), while the branch 5 was withdrawn by irradiation as its phase difference wildly fluctuated. In general, such an antiphasic synchronization of r_i(t) and r_c(t) is notable when the branch i grows rapidly, whereas in-phase synchronization (d_i(t) ≃ 0°) cannot yield rapid growth. Although growth movements are inhibited by irradiation in principle, it was confirmed that even under irradiation, the rapid growth rate is as high as that under nonirradiated conditions when exact antiphasic synchronization occurs (data not shown). This is because protoplasmic streaming in-
ducing the branch growth is driven by the pressure difference (gradient) derived from the contraction tension difference between the center and the periphery of the gel layer. Namely, an antiphasic branch grows while its relaxing gel layer invites a protoplasmic influx as the contracting gel layer in the center disc extrudes the protoplasm. Because the positioning of the antiphasic branches is essential for the organism’s decision on its shape deformation, we refer to the spatial arrangement of branches sustaining antiphasic synchronization for several periods as “the oscillation pattern” for simplicity in the following discussions.

To destabilize a once-recalled memory configuration, it is necessary that at least one branch newly emerges into an irradiated neuron and sustains its growth for sufficiently long periods. Usually a new branch emerges suddenly as a lumplike bulge structure with strikingly large amplitude of thickness oscillation, by intensively inviting a protoplasmic influx from a broad domain (Figs. 2C and E). Our speculation is that the bulge is developed as protoplasmic sol spouts from a local site of the gel layer where it is sufficiently fragile to be ruptured by hydrostatic pressure of protoplasm, as similar phenomena are observed when we scratch on the gel layer to create a local defect artificially.

Even if a new branch (bulge) emerges, whether the branch can sustain its growth depends on how the global body self-organizes the oscillation pattern. Protoplasm supply for sustained growth of a branch is reduced if it disperses into other antiphasic branches. Conversely, as shown in Fig. 3, the invading branch 4 could sustain its growth as the dispersive oscillation pattern (left) evolved into the proper oscillation pattern (middle) such that protoplasm in the irradiated branch 5 was exclusively supplied to the antiphasic branch 4.

As well as the results reported by Takamatsu [2], spontaneous transition of the oscillation pattern could be observed even when the organism’s planar shape is almost unchanged. The transition has a stochastic nature as a pattern can evolve into various patterns. This stochastic transition makes it possible to eventually attain the proper oscillation pattern capable of sustaining the growth of newly-emerged branches, and has a functional significance in switching between the stabilizing and destabilizing modes.

In summary, we highlighted two key factors to yield the memory destabilization; the transition of the oscillation pattern and the emergence of a new branch as a bulge structure. As for the former, we suppose that there is some local-to-global fluctuation amplification process. Namely, local fluctuations in the organism’s internal conditions, such as the thermal fluctuation and/or nonuniformity of chemical distribution, are amplified to an extensive and sustained fluctuation through some form of positive feedback effect by the coupling of physiological and hydrodynamic processes, and eventually influence the global oscillation pattern. As for the latter, we should examine the validity of the global-to-local effect, such as an influence of the global oscillation pattern on creating a local defect of the gel layer from which a bulge develops, because the gel layer may be ruptured when the local site is intensively pressured by a broad domain. The organism’s survival abilities in a harsh environment, therefore, may be produced by the circulation of the local-to-global and global-to-local effects [5].

5. Concluding Remarks

In this study, we experimentally implemented an amoeba-based neurocomputer as an associative memory device capable of the dynamic memory transition. Our system can be developed to an optimization problem solver by properly altering the number of neurons and the assignment of weights as Hopfield and Tank originally proposed their model in a continuous form to obtain a valid solution of the traveling salesman problem [4]. In either case, memories or valid solutions are represented as stable minimum-energy states of a potential landscape established by the weights. In the previous work, the amoeboid organism’s capacity of maze solving, which is a kind of optimization problem solving, was shown by focusing on a stabilizing process to converge toward the shortest route solution that optimizes mainly the chemical transport efficiency on a physical metric space [1]. In contrast, we observed that the organism’s dynamics produces not only stabilization but also destabilization without explicitly applying external perturbation, and that the spontaneous switching between the stabilizing and destabilizing modes may assist searching for a global optimum solution without being anchored at local optima. Furthermore, our result suggests that the organism may be capable of solving optimization problems even on a more abstract space than the physical metric space, a phase space of the oscillation patterns, where a solution is obtained by optimizing spatial arrangements of temporal oscillation phases. It will provide us a viewpoint of discussing the functionalities of the unicellular amoeboid organism’s oscillatory behavior in analogy with the brain activities.

References


766
Synchronization Experimentally Observed in Calling Behaviors of Japanese Rain Frogs (Hyla-japonica)

Ikkyu Aihara\(^1\), Hiroyuki Kitahata\(^2\), Shunsuke Horai\(^4\), Kazuyuki Aihara\(^3\), and Kenichi Yoshikawa\(^2\)

\(^1\) Department of Applied Analysis and Complex Dynamical Systems, Graduate School of Informatics, Kyoto University
\(^2\) Department of Physics, Graduate School of Science, Kyoto University
\(^3\) Institute of Industrial Science, University of Tokyo
\(^4\) Aihara Complexity Modelling Project, ERATO, JST

Email: aihara.ikkyu@t02.mbox.media.kyoto-u.ac.jp

Abstract — We recorded calls of Japanese rain frogs (Hyla-japonica) and analyzed the nonlinear dynamics on the basis of the experimentally observed data. The result shows that a single frog called nearly periodically, and a pair of frogs called alternately in almost anti-phase synchronization through mutual information. We also studied the behaviors with mathematical models.

1. Introduction

Synchronization is everywhere in coupled nonlinear systems with oscillations [1–3]. In particular, synchronization of biological oscillators has been studied extensively and intensively [3,4]. In this study, we analyse calling behaviors of frogs. There exist some studies on synchronization of calls of frogs. Loftus-Hills studied the synchronization in calling behaviors of frogs Pseudacris streckeri [5], while Lemon and Struger studied acoustic entrainment to randomly generated calls in frogs Hyla crucifer [6]. In the present paper, we study calling behaviors of Japanese rain frogs Hyla japonica [7].

2. Methods

Male Japanese rain frogs Hyla japonica were collected in paddy fields in Kyoto, Japan (see Figs. 1 and 2). Collected frogs were individually housed in small plastic cages which were dipped about 1cm in water of the paddy fields where the frogs inhabited, as shown in Fig. 3.

Time series data of calls were examined with nonlinear time series analysis [8–10] including the recurrence plot (RP) [11] and its extensions [12], namely the iso-directional recurrence plot (IDRP) and the iso-directional neighbors plot (IDNP). The RP is a two-dimensional representation which plots point \((i,j)\) whenever the distance between the points \(X_i\) and \(X_j\) on a reconstructed orbits is less than or equal to the threshold \(r\): \(||X_i - X_j|| \leq r\). Similarly to RP, the IDRP characterizes the movement of orbits. That is, point \((i,j)\) is plotted if \(||(X_{i+T} - X_i) - (X_{j+T} - X_j)|| \leq s\), where \(s\) is the threshold, \(T\) is the delay for IDRP which equals to the reconstruction delay, and the thresholds with \(s = r\) are defined so that
Figure 3: Plastic cages to house frogs.

10% of the recurrences are plotted in each picture. The IDNP is simply the product of RP and IDRП: $IDNP = RP \cap IDRП$. Therefore, IDNP represents the points that are close to each other and are moving in a similar direction in the reconstructed state space, which is a characteristic peculiar to deterministic dynamics.

3. Experimental Results

Figure 4 shows the time series data of calls recorded from (a) a single frog and (b) two mutually interacting frogs, where the upper and lower panels in (a) and (b) show the waveforms and the spectrograms in time windows of 1.8s, respectively. A single frog called nearly periodically as shown in Fig. 4(a). In the experiment shown in Fig. 4 (b), after one frog called several times alone, the second frog began to call as the timing is shown by the red arrow, and two frogs called alternately in almost anti-phase synchronization.

Figure 5 shows reconstructed orbits of the calls of (a) a single frog and (b) two frogs in delay coordinates [8–10]. The orbits of Fig. 5 are reconstructed from the time series data of amplitudes of the calls which are re-bined with 500 points, where the total time length is 3.8s. Figure 5 shows that the calls of both a single frog and two frogs are nearly periodic; the latter has roughly two-folded structure. The power spectrum analysis showed that the fundamental frequency of calling by a single frog in Figs. 4(a) and 5(a) and that of calling by two interacting frogs in Figs. 4(b) and 5(b) are 4.2Hz and 7.0Hz, respectively [7]. The latter value means that the first frog was calling with 3.5Hz in the mutually interacting behavior. In other words, the fundamental frequency of the calls of a single frog during the interactive calls between two frogs is smaller than that of the same frog which called alone. Figures 6(a), (b), and (c) show the RP, the IDRП, and the IDNP of the interacting calling data. These plots suggest existence of nonlinear deterministic dynamics with emergence, disappearance, and re-emergence of anti-phasely synchronized calling.

4. Mathematical Analysis

Synchronization of biological oscillators has been theoretically analyzed [1–4,13,14]. Based on these previous studies, we model the calling behavior of mutually interacting frogs with two mutually coupled phase oscillators as follows [7]:

$$\frac{d\theta_a}{dt} = \omega_a - g_{ab}(\theta_b - \theta_a - \alpha), \quad (1)$$

$$\frac{d\theta_b}{dt} = \omega_b - g_{ba}(\theta_a - \theta_b - \beta), \quad (2)$$

where $\theta_a \in S^1$, $\theta_b \in S^1$, $S^1 = [-\pi, \pi]/(\pi \equiv -\pi)$ [14], $\omega_a$ and $\omega_b$ are the natural calling frequencies of two frogs, $g_{ab}$ and $g_{ba}$ are $2\pi$-periodic functions that represent the mutual interactions, $\alpha$ and $\beta$ are positive frustration parameters. Subtracting Eq.(2) from Eq. (1) gives

$$\frac{d\phi}{dt} = (\omega_a - \omega_b) + g_{ba}(\phi - \beta) - g_{ab}(-\phi - \alpha), \quad (3)$$

where $\phi$ is the phase difference with $\phi = \theta_a - \theta_b$.

If we assume for the sake of simplicity that $g_{ab}(\psi) = g_{ba}(\psi) \equiv K \sin(\psi)$ [1,2], then, Eq. (3) can be represented as follows:

$$\frac{d\phi}{dt} = (\omega_a - \omega_b) + 2K \cos(\frac{\alpha + \beta}{2}) \sin(\phi + \frac{\alpha - \beta}{2}). \quad (4)$$

Under the assumptions of $|\omega_a - \omega_b| \ll K$, $\alpha \approx \beta$, and $\alpha \ll \pi$, the stable equilibrium point $\phi^*$ nearly equals to $\pi$, which corresponds to almost anti-phase synchronization. This result represents the property that two frogs call alternately [7].
Figure 4: The time series data of calls of (a) a single frog and (b) two mutually interacting frogs. The upper and lower panels in (a) and (b) show the waveforms and the spectrograms in time windows of 1.8s, respectively. The red arrow in Fig. 4(b) shows the timing when the second frog began to call.

Figure 5: Reconstructed orbits in delay coordinates of the calls of (a) a single frog and (b) two frogs, where the delay is set at 0.015s. The total time length is 3.8s both in (a) and (b).

Figure 6: (a) the RP [11], (b) the IDRP, and (c) the IDNP [12] of the interactive calling data with the embedding dimension 5 and the delay time 0.03s, where the original data are squared and averaged over a window of 0.05s, and re-sampled at each 0.01s.
5. Conclusion

We recorded and analyzed the time series data of calls of Japanese rain frogs. The result shows that a single frog called nearly periodically, and a pair of frogs called alternately; after one frog called several times alone, the second frog began to call, and two frogs called alternately in almost anti-phase synchronization. The frequency of the calls of the first frog during the interactive calls of two frogs is smaller than that of the same frog which called alone. We modeled these phenomena as the system of coupled phase oscillators which synchronize in almost anti-phase. The experimentally observed anti-phase synchronization implies a possibility that two male frogs *Hyla japonica* call alternately in order to make female frogs distinguish them each other [7].

6. Acknowledgments

This study was partially supported by a Grant-in-aid for the Scientific Research (No.18GS0421) “Construction of functional de novo cell” and the 21st Century COE Program “Center of Excellence for Research and Education on Complex Functional Mechanical Systems” at Kyoto University from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

We’d like to thank Y. Kuramoto, I. Tsuda, H. Kori, H. Fujisaka, A. Nagano, and K. Sogo for their stimulating discussions and valuable comments.

References


Measurement noise reduction by local projection methods

M. E. Mera and M. Morán
Departamento de Análisis Económico I, Universidad Complutense
Campus de Somosaguas, 28223 Madrid, Spain
Email: mera@csee.ucm.es, mmoranca@csee.ucm.es

Abstract—A measurement noise reduction problem occurs when a sequence of states (time series) of a system governed by a low dimensional dynamics is recorded using a measurement process subject to error. In this paper we review those algorithms designed for noise reduction with unknown dynamics and that take a non-parametric approach, i.e. such that the only assumption on the underlying dynamics is its smoothness. These algorithms are always based on the local analysis of the data, which are a scalar time series in some cases and a multivariate time series in other cases. We show that all of the algorithms can be understood in a common framework: the projection of neighbourhoods on suitable linear manifolds. In this context the essential distinctive features of each algorithm may be explained in terms of the metric considered, the neighbourhouds chosen, and the goal pursued.

1. Introduction

Many problems relating to measurement noise reduction can be described by the equations

\[ x_{k+1} = f(x_k) \text{ and } X_k = g(x_k) + e_k. \]

where \( x_k \in \mathbb{R}^d \) is the unobservable state vector of a system at period \( k \); \( f \) is a smooth unknown dynamics, \( g : \mathbb{R}^d \to \mathbb{R}^p \) is a smooth vectorial function or observable; \( e_k \) is an independent and identically distributed (i.i.d.) multivariate stochastic process; and \( X_k, k = 1, 2, ..., N \), is the available time series. In this scheme \( e_k \) is referred to as the measurement noise. Further, we adopt a non-parametric approach, which does not assume any specific form for the dynamics. Instead, it is based on a local analysis which tries to exploit the smoothness of \( f \). There is a rich literature on algorithms that fall into this category (see [1]-[8]). They can be classified as those algorithms for which the function \( g \) is the identity and those designed for scalar time series, i.e. \( g : \mathbb{R}^d \to \mathbb{R} \). These last algorithms require the reconstruction of a multidimensional state space equivalent in some sense to the original one. Takens’s theorem guarantees [9] that in absence of noise, for generic observable \( g \) and if \( m \geq 2d + 1 \), the \( m \)-dimensional reconstructed time series \( x^m_k := (x_k, x_{k+1}, ..., x_{k+m-1}) \), \( k = 1, 2, ..., N-m+1 \), provides a diffeomorphic image of the attractor associated with \( f \) and that the main properties of the dynamics \( f \) can be inferred from those of the shift dynamics \( f'(x^m_k) := x^m_{k+1} \).

One difference between noise reduction for scalar and for multivariate time series is the structure of the error term. In scalar time series the error is unidimensional, so the degrees of uncertainty of the \( m \) coordinates of the reconstructed time series \( x^m_k := x^m_k + e^m_k \) are identical, i.e., \( e^m_k \) is an uncorrelated and homoskedastic error (it has covariance matrix \( \Sigma = \sigma^2 I \) where \( I \) is the identity matrix). The algorithms designed for scalar time series can be used for multivariate time series corrupted by uncorrelated and homoskedastic noise. However, the results can be improved significantly [10] if we incorporate into the reduction scheme information about the structure of the error term when it is correlated and/or heteroskedastic.

All of the algorithms we describe in this paper are based on local projections. The idea is that in absence of noise the reconstructed time series \( x^m_k, k = 1, ..., N - m + 1 \) lies on a \( d \)-dimensional smooth submanifold \( M \) of \( \mathbb{R}^m \). Therefore, for points \( x^m \) within a small ball \( U_i \) centered at \( x^m \), the \( m \)-dimensional difference vectors \( x^m - (x^m)_{U_i} \), where \( (x^m)_{U_i} \) is the center of mass of \( U_i \), span only a \( d \)-dimensional linear subspace, which is in fact an estimate of the tangent space of \( M \) at \( (x^m)_{U_i} \). However, these difference vectors span \( \mathbb{R}^m \) due to the noise, so the deterministic part of the time series can be (partially) recovered by projecting such vectors onto suitable \( d \)-dimensional linear subspaces.

For multivariate \( d \)-dimensional time series the algorithms first proceed to embed the time series in a higher dimensional space \([3], [10]\), considering for instance \( Z_i = (X_{i-1}, X_i, X_{i+1}), i = 2, ..., N - 1 \) or \( Z_i = (X_i, X_{i+1}), i = 1, ..., N - 1 \). Since in absence of noise, these data lie also on a \( d \)-dimensional smooth submanifold, the algorithms recover the deterministic part of the time series projecting again on suitable \( d \)-dimensional subspaces.

Therefore, the first distinctive feature of each (scalar or vectorial) algorithm is what embedding space is to be used (i.e. the working space). This determines what neighbourhoods are to be chosen. The second feature is the metric to be considered in the embedding space. We show below how the estimations given by each algorithm can be understood as a local projection with respect to a given metric.

The third distinctive feature of the algorithms is their goal. All of the algorithms are iterative: the noise reduction scheme produces a new time series that will be the input for the following iteration of the algorithm. In order to decide the end of the process we need to measure the
level of reduction of noise. Some authors [1]-[3], [6]-[8] try to reduce the pointwise distance between the cleaned time series \( \tilde{x}_i \), \( i = 1, \ldots, N \) and the clean one, 

\[
E(\mathbf{x}, \tilde{\mathbf{x}}) := \left( \frac{1}{N} \sum_{i=1}^{N} \| \tilde{x}_i - x_i \|^2 \right)^{1/2}.
\]

This is consistent with the goal of recovering the original, uncorrupted time series. Other authors [5] try to maximize the local self-consistency of the data by reducing

\[
E_{dyn}(\mathbf{x}) := \left( \frac{1}{N} \sum_{i=1}^{N} \min_{j=1,\ldots,N} \| \tilde{x}_i - \tilde{j}(\mathbf{x}_i) \| \right)^{1/2},
\]

(1)

where \( \tilde{j}_i \) is a local estimation of the dynamics \( f \) at \( x_i \). This provides a dynamical noise reduction.

Mera and Morán have recently proposed an alternative approach [10]. They try to recover the geometric properties and the long-run statistical regularity of the underlying dynamics by minimizing the mean distance to the attractor

\[
d_{m}(\mathbf{x}, \mathbf{x}) := \frac{1}{N} \sum_{i=1}^{N} \min_{j=1,\ldots,N} \| \tilde{x}_i - \tilde{j}(\mathbf{x}_i) \|.
\]

Consider that any true orbit of the dynamical system converges (with probability one) rich information on the geometry of the attractor and on the statistical long-run behaviour of the dynamics \( f \). The \( d_{m} \) distance between any such orbit and the original clean orbit is very small. In contrast, the pointwise distance between both orbits is, in general, large. The noise reduction measure associated with \( d_{m} \) is \( R_m := 100 (1 - d_{m}(\mathbf{x})/d_{m}(\mathbf{x})) \). We shall refer to this noise reduction measure as statistical noise reduction.

Creating a common framework for the different noise reduction algorithms considered here is made possible by the following theorem, which shows how to get the linear subspace which best fits data in \( \mathbb{R}^n \) with respect to a given metric. Let \( A \) be an \( n \times n \) symmetric positive definite matrix and let \( L_p \) be the set of \( p \)-dimensional linear subspaces of \( \mathbb{R}^n \). The orthogonal projection of \( \mathbf{v} \in \mathbb{R}^n \) onto \( T \in L_p \) with respect to the metric \( d_{m} \) is given by

\[
P_T \mathbf{v} := \arg \min_{\mathbf{w} \in \mathbb{R}^n} \| \mathbf{v} - \mathbf{w} \|^2.
\]

We say that the linear subspace \( T_p \) is the best linear subspace in \( L_p \), for the points \( \{Z_k, k = 1,\ldots,N\} \in \mathbb{R}^n \), with respect to the metric \( \delta_{p} \), if it minimizes over all \( T \in L_p \)

\[
E(T) := \frac{1}{N} \sum_{k=1}^{N} \| (Z_k - P_T Z_k) A^{-1} (Z_k - P_T Z_k) \|.
\]

Such best linear subspaces can be obtained by the eigenvectors of the \( n \times n \) matrix

\[
M_{ZZ} := \frac{1}{N} \sum_{i=1}^{n} Z_i Z_i^T
\]

in the metric \( \delta_{p} \). A system of vectors \( \{w_i, i = 1,\ldots,n\} \in \mathbb{R}^n \) is called orthonormal system of eigenvectors of \( M_{ZZ} \) in the metric \( \delta_{p} \) if they satisfy: (i) there exist real numbers \( \lambda_i \) (called eigenvalues in the metric \( \delta_{p} \)) such that \( M_{ZZ} \mathbf{w}_i = \lambda_i \mathbf{w}_i \), \( 1 \leq i \leq n \), and (ii) \( \mathbf{w}_i \mathbf{w}_j = \delta_{ij} \), for all \( i, j \).

**Theorem** (see [10]) Let \( \lambda_1 \geq \cdots \geq \lambda_n \) be the eigenvalues of \( M_{ZZ} \) in the metric \( \delta_{p} \), let \( \{w_i, i = 1,\ldots,n\} \) be the corresponding orthonormal eigenvectors and, for \( 1 \leq p \leq n \),

let \( B \) be the \( n \times (n - p) \) matrix whose columns are the last \( n - p \) eigenvectors \( \{w_{p+1},\ldots,w_n\} \). Then the best linear subspace in \( L_p \) with respect to the metric \( \delta_{p} \) is given by \( T_p := \text{span} \{Aw_1,\ldots,Aw_p\} \) and \( P_{T_p} Z := (I - ABB^T)Z, Z \in \mathbb{R}^n \).

2. Algorithms for scalar time series


These algorithms use as working space \( \mathbb{R}^m \) for some suitable \( m \), and they consider the Euclidean metric, i.e. they take the \( m \)-dimensional identity matrix \( I \) as the matrix \( A \) in the above theorem. Let \( U_i \) be a neighbourhood of \( x_i \) and let \( \{X_{i}^m\}_{i=1}^{N} \) be the center of mass of the points within \( U_i \). The algorithms reduce the noise by projecting the data \( Z_j := X_j^m - \{X_{i}^m\}_{i=1}^{N} \), \( X_j^m \in U_i \) onto the best \( d \)-dimensional subspace \( T_d \). The estimation of \( x_j \) they compute is \( \tilde{x}_j = \{X_{i}^m\}_{i=1}^{N} + \alpha Z_j + (1-\alpha)P_{T_d} Z_j \) where \( 0 \leq \alpha < 1 \). Notice that this procedure ties up the cleaned time series to the original data through the term \( \alpha Z_j \). This is in accordance with the goal of a pointwise reduction. In fact, the authors justify this step as a useful means to soften the effects of rare statistical outliers. Since each point of the scalar time series appears as a component of \( m \) consecutive delay vectors, the algorithm provides multiple estimations for almost all of the points of the scalar time series. This problem is settled by averaging such estimations.

2.2. Schreiber-Grassberger method [8]

After a previous embedding of the data in \( \mathbb{R}^m \) for a suitable (odd) value of \( m > 2d + 1 \), this algorithm proceeds to assign as the estimate of the central coordinate of each \( m \)-vector \( X_i^m \) a linear function of its \( m - 1 \) remaining coordinates. Such linear function is estimated using a neighbourhood \( U_i \) of \( X_i^m \). This gives a single estimate of each data point of the scalar time series. It can be proved that this procedure is equivalent to the projection of \( X_i^m \) on the best \( m - 1 \) dimensional subspace with respect to the metric \( \delta_{p} \) corresponding to a diagonal matrix \( A \) with all its entries equal to zero, with the exception of the central coordinate which is set equal to one. Since such a matrix is not invertible, the authors take a diagonal matrix whose diagonal entries are all very small (in the numerical results they take a value 0.001) and a single 1 in the central entry of the diagonal. Here the central coordinate plays an special role: If the coordinate computed as a linear function were to be the last coordinate, it would be ill-defined along the unstable manifold and a correction of the first coordinate would be ill-defined along the stable manifold. Only the central coordinate correctly takes into account information about both the past and the future. It should be mentioned here the very simple but quite efficient noise reduction algorithm [7] which proceeds in the same way as described above, but takes as an estimate of the central coordinate the
average of the central coordinates for all the points in the
neighbourhood.

2.3. Grassberger et al. method [2]

This algorithm is a modification of the previous one but
instead of correcting just the central coordinate for
each point \( X_i \), it corrects several more central coordinates.
How many of these will be corrected and the dimension of
the projection subspaces are parameters of the algorithm.
Since each data point of the scalar time series appears as a
one central component of several delay vectors, the average
of such estimations gives the final estimate of the point.
In our setting, the metric \( \delta_A^{-1} \) is given by a diagonal matrix \( A \)
all with very small entries, with the exception of the central
coordinates which are equal to ones.

3. Algorithms for multivariate time series


The smoothness of \( f \) implies that, and for any point
\( x \) in a small neighbourhood of a point \( x_i \) in the attractor
\( f(\mathbf{x}) \sim \mathbf{B} x + \mathbf{d} \), holds, where \( \mathbf{B} \) is some \( d \times d \)
matrix and \( \mathbf{d} \) is a \( m \)-dimensional vector. Thus, the algorithm
takes a neighbourhood \( U_i \) for each point \( x_i \) of the noisy \( d \)
dimensional time series, and then gives as estimates of \( \mathbf{B} \)
and \( \mathbf{d} \) the solution to the least squares problem

\[
\min_{\mathbf{B}, \mathbf{d}} \sum_{x_i \in U_i} \| \mathbf{x}_{j+1} - (\mathbf{B} \mathbf{x}_j + \mathbf{d}) \|^2.
\]

A naive estimate \( \mathbf{x}_i \) of \( x_i \) would be \( \mathbf{x}_i := \mathbf{B} \mathbf{x}_i + \mathbf{d}_i \), but such
an estimate is too influenced by the noise in \( \mathbf{x}_i \) and by
the error in the estimation of \( \mathbf{B} \) and \( \mathbf{d} \). For this reason the
algorithm seeks a new time series which is consistent with
the estimated linear maps. This is done by selecting windows
of \( p \) consecutive points \( \{ \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+p} \} \) and then
taking \( \{ \mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+p} \} \) as the solution of the optimization problem

\[
\min_{\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+p}} \sum_{j=1}^{p} \| \mathbf{x}_i - \mathbf{x}_j \|^2 + \sum_{j=1}^{p} \| \mathbf{x}_{j+1} - (\mathbf{B} \mathbf{x}_j + \mathbf{d}_j) \|^2 \tag{2}
\]

where \( w \) is a weighting factor. Notice that this minimization
problem is coherent with a dynamical noise reduction
(see (1)). The link with the projection methods comes from
the fact that the least squares estimator is also a projection
matrix.

3.2. Hegger and Schreiber’s method[3]

This algorithm is conceptually similar to the previous
one. It is also an adaptation to multivariate time series
of Schreiber and Grassberger’s algorithm [8] described in
the previous section. In order to capture the information
about the past and the future in the estimate of \( x_i \) the
algorithm takes the time series \( \mathbf{Z}_i := \{ \mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1} \} \in \mathbb{R}^d \),
for \( i = 2, \ldots, N - 1 \). Thus the working space is \( \mathbb{R}^d \), with \( d \)
the dimension of the original data points. The hypothesis
is that the clean time series satisfies, for points \( z_j \) near to
\( z_i := (x_{i-1}, x_i, x_{i+1}) \), the linear relationship

\[
x_j \sim \mathbf{B} x_{j-1} + \mathbf{C} x_{j+1} + \mathbf{d}_j
\]

where \( \mathbf{B} \) and \( \mathbf{C} \) are \( d \times d \) matrices and \( \mathbf{d}_j \) is an \( d \)-
dimensional vector. They are computed by solving the optimization
problem

\[
\min_{\mathbf{B}, \mathbf{C}, \mathbf{d}} \sum_{x_i \in U_i} \| \mathbf{x}_j - (\mathbf{B} \mathbf{x}_{j-1} + \mathbf{C} \mathbf{x}_{j+1} + \mathbf{d}_j) \|^2
\]

where \( U_i \) is a neighbourhood of \( Z_i \). Let \( Z_{j \text{corr}} := \mathbf{B} x_{j-1} + \mathbf{C} x_{j+1} + \mathbf{d}_j \) be the estimate given by the linear model
described above. The algorithm takes as estimate \( \hat{Z}_i = (1 - \alpha) \hat{Z}_i + \alpha Z_{j \text{corr}} \) where \( 0 < \alpha \leq 1 \). It can be proved that
\( Z_{j \text{corr}} \) are the central coordinates of the orthogonal projection
of \( Z_i \) on the best linear \( d \)-dimensional subspace with
respect to the metric \( \delta_A^{-1} \), where \( A \) is a diagonal matrix having
all the entries almost null except for the \( d \) central coordinates,
which are ones.

3.3. Mera and Morán’s method[10]

This algorithm pursues a statistical noise reduction, and
it is the only algorithm in the literature able to exploit
the available information that an uncorrelated and heter-
skedicastic noise contains. Its starting point is Kostelich
and Yorke’s algorithm [5]. In fact these authors already
pointed out the unsuitability of the least squares fits.
Indeed it is well known that in linear models where the
independent variables are also measured with error, the least
squares estimators are biased, being the bias an increasing
for these type of models showing how to get unbiased and
consistent estimators, which are also the maximum likeli-
hood estimators if the errors are Gaussian.

Mera and Moran’s algorithm uses \( \mathbb{R}^d \) as working space,
and \( d \) the dimension of the data. It starts by taking a neigh-
borhood \( U_i \) for each point \( x_i \) and it uses the theory devel-
one in [11] to obtain an unbiased maximum likelihood esti-
mate of the matrix \( \mathbf{B} \), satisfying

\[
\mathbf{X}_{j+1} - \langle \mathbf{X}_{i+1} \rangle_{U_i} \approx \mathbf{B} \langle \mathbf{X}_j - \langle \mathbf{X}_i \rangle_{U_i} \rangle_{U_i}, \mathbf{X}_j \in U_i
\]

where \( \langle \mathbf{X}_i \rangle_{U_i} \) denotes the center of mass of the points of
the neighborhood \( U_i \) and \( \langle \mathbf{X}_{i+1} \rangle_{U_i} \) is the center of mass
of their images. Furthermore, this theory gives, for each
\( \mathbf{X}_j \in U_i \), the maximum likelihood estimate \( \hat{Z}_j \) of \( z_j := (\mathbf{x}_{j-1} - \langle \mathbf{x}_{i+1} \rangle_{U_i}, \mathbf{x}_j - \langle \mathbf{x}_i \rangle_{U_i}) \).
Using such estimation makes unnecessary the last step in the Kostelich
and Yorke’s algorithm (see (2)).

The link between Mera and Moran’s algorithm and the
projection methods comes from the fact that \( \hat{Z}_i \) coincides
[10] with the orthogonal projection of \( Z_j \) on the best \( d \)
dimensional linear subspaces in the metric \( \Sigma^{-1} \), where \( \Sigma \)
is the covariance matrix of the errors \( e_j \), in \( \mathbf{Z}_j := z_j + e_j \),
Thus, this method takes advantage of the structure of Σ to reduce the noise. The authors prove that this method gives a time series having almost the same statistical and geometric properties as the true dynamics, even in the case of high noise amplitudes or when the variances of the components of the error are different and/or there are correlations. Since, in general, the matrix Σ is unknown, the identity matrix is taken as Σ at the first iteration of the algorithm, and it is updated in the next iterations with the empirical covariance matrix of the estimated errors.

The figure above shows a noisy time series of 10000 points generated by the Hénon map, with a highly heteroskedastic Gaussian noise (σ_ex = 0.01σ_s and σ_ey = 0.15σ_s, where σ_s is the standard deviation of the clean signal). We have plotted the output of the algorithm together with the noisy time series. The levels of reduction are 83% in terms of pointwise distance, 97% in terms of dynamic consistency and 94% in terms of statistical regularity.

4. Conclusions

We have reviewed several algorithms addressed to measurement noise reduction and based on a non-parametric, local approach. All these algorithms can be understood as producing outputs through local orthogonal projections, with respect to suitable metrics, onto linear manifolds. They differ in the working spaces, the metrics and the goals pursued. Although these are the three essential distinctive features, there are also other special useful devices proposed by the authors of the algorithms whose discussion is beyond the scope of this short review. Interested readers can find these details in the quoted literature. In particular criteria are available for an adaptive selection of the size of the neighbourhoods [13], and for a proper selection of the dimension of the subspace in which the neighbourhoods must be projected [12], both depending on the geometry of the time series at each point instead of remaining fixed for all the points.

All of the algorithms display a highly efficient performance in their respective goals. For instance for a time series of 5000 data points from Hénon map corrupted with an uncorrelated and heteroskedastic Gaussian noises with amplitudes between 1% and 10% they give a noise reduction in terms of pointwise distance between 65% and 73% and between 83% and 87% in terms of dynamical consistency. These results seem to indicate that, in spite of the stronger intuitive content of pointwise noise reduction and the appealing fact that pointwise noise reduction permits the improvement of the short run forecasting of the series, there exist intrinsic bounds for such goal. These bounds might be related to our limited capability for a pointwise, short run prediction of chaotic dynamical systems recorded with uncertainty, even if the dynamics is known. By contrast, the more realistic goal of statistical noise reduction could find its bounds only in the length of the data sets.

Acknowledgments

We would like to thank NOLTA2006 organizing committee members for their fruitful suggestions and comments. This research has been supported by the Ministerio de Ciencia y Tecnología, research project BFM2003-08204/Mate.

References

A coalgebraic approach to behavioral inheritance in Dynamical Systems

Cristian Secchi†, Marcello Bonfè‡ and Cesare Fantuzzi‡

†DISMI, Università di Modena e Reggio Emilia, Italy
Email: {secchi.cristian, fantuzzi.cesare}@unimore.it
‡ENDIF, Università di Ferrara, Italy
Email: mbonfe@ing.unife.it

Abstract—The paper introduces a formal definition of inheritance, which is a cardinal concept in object-oriented software, for dynamical systems. The proposed definition exploits the coalgebraic description of software artifacts to provide a connection between the behavioral approach for modeling dynamical systems and the object-oriented approach for software modeling and design.

1. Introduction

The object-oriented (o-o) approach is renowned as a driving idea for many modern software design methods, supported by several programming languages of common use. One of the reasons for the success of object-orientation is the fact that it allows to design by extension, exploiting a feature called inheritance. Thanks to the inheritance mechanism, given a basic class of software objects, characterized by a set of attributes (data) and methods (functions), it is possible to build a new class which inherits all the characteristics (i.e. attributes and methods) of the basic one and which can be further endowed with additional attributes and methods. However, inheritance, may also become an issue in terms of behavioral conformity between basic and derived classes. Theoretically, the Liskov Substitution Principle (LSP) [3] states that an object of a derived class can be asked to do anything an object of the basic class can do, which means conformity of functional interface, but also that it should do it in the very same way and with the same results, which means conformity of behavior. If we consider object-orientation as a conceptual approach to network-based modeling of dynamical and physical systems, as supported by languages like Bond-Graphs [2], Modelica [6] and the extension of UML defined in [5], or to the design of mechatronic systems, the advantage of inheritance is evident: a complex system composed of a network of, possibly nonlinear, systems processing physical information (i.e. energy) can be extended by replacing one of its parts with a derived component which is able to realize the basic behavior, but, in addition, also to perform further functionalities. Clearly, in this domain it is necessary to define inheritance in a way that guarantees behavioral conformity.

The present paper presents an approach to formalize inheritance in an o-o framework for dynamical systems modeling. The major contribution of the proposed approach consists in allowing designers of mechatronic systems to adopt the principle of design by extension, so that they can formally specify new mechatronic components with additional features and use them to replace parts of a more complex system, without modifying its global behavior. The o-o framework presented in the paper combines the coalgebraic view of object-orientation presented in [1] and the behavioral approach [4]. Next section will give a brief overview of basic ideas of the two motivating references, while Sec. 3 describes the proposed o-o approach to Dynamical Systems modeling and the definition of inheritance for this domain. The paper ends with some ideas for future enhancements and applications of the proposed modeling framework.

2. Background

The behavioral approach: this mathematical framework aims to formalize a general dynamical system by focusing on its behavior, rather than on the equations that characterize it. For an extensive description of this approach see [4]. Consider a dynamical system characterized by an I/O interface $(u, y) \in U \times Y$ through which it can interact with the rest of the world. The interface outcomes space, namely the space on which inputs and outputs of the system live, is defined as $W = U \times Y$. It is then possible to define the universe of interface outcomes $U$ as the set of ALL possible trajectories that can be described in $W$, namely as

$$U = \{(u, y) \in W^T\}$$  \hspace{1cm} (1)

where $W^T$ is the set of maps from $T$, an ordered time set, to $W$. An interface behavior is defined as a certain $B \subseteq U$ of compatible interface outcomes, namely a particular subset of trajectories that the system is allowed to describe in $W$. The interface behavior $B$
can be described as the set of possible solutions of a group of differential-algebraic equations, called behavioral equations: the interface outcomes of $B$ are all the pairs $(u(t), y(t))$ that satisfy the behavioral equations; for further information see [7].

The coalgebraic modeling framework: category theory and coalgebras are the basis of a modeling framework for o-o programming developed in [1]. Category theory is a branch of mathematics that has been developed to describe various structural concepts from different mathematical fields in a uniform way. Intuitively category theory discusses a class of objects and their relationship to each other and it is formally defined as follows:

**Definition 1** A category consists of

- A collection of objects $(A, B, C, \ldots)$
- A collection of morphisms $(f, g, h, \ldots)$, each one relating two objects (usually written $f : A \to B$).
- Given morphisms $f : A \to B$ and $g : B \to C$ there is a morphism $g \circ f : A \to C$ called the composite of $f$ and $g$.
- For each object $A$ there is a morphism $1_A : A \to A$ called the identity morphism of $A$.

These data are required to satisfy the associativity law (i.e. $h \circ (g \circ f) = (h \circ g) \circ f$ for all $f : A \to B$, $g : B \to C$, $h : C \to D$) and the unit law (i.e. $f \circ 1_A = f = 1_B \circ f$ for all $f : A \to B$).

A functor is a mapping from one category to another that preserves the categorical structure; it maps objects into objects and morphisms into morphisms and it preserves the associativity and the unit property. A functor that maps a category into itself is called endofunctor. Let $\mathfrak{A}$ and $\mathfrak{B}$ be categories such that each object $A \in \mathfrak{A}$ can be regarded as an object of $\mathfrak{B}$ by suitably ignoring structures that $A$ may have as an object of $\mathfrak{A}$ but not as an object of $\mathfrak{B}$. A functor $\mathcal{F} : \mathfrak{A} \to \mathfrak{B}$ which operates on objects of $\mathfrak{A}$ by “forgetting” any imposed mathematical structure is called a forgetful functor. Loosely speaking a forgetful functor is a kind of generalized canonical projection which cuts away all the information which does not fit in the target space. Another important definition in the coalgebraic framework is, of course, the one of coalgebra:

**Definition 2** Let $\mathfrak{A}$ be an arbitrary category and let $\mathcal{F} : \mathfrak{A} \to \mathfrak{B}$ be an endofunctor. A coalgebra consists of an object $X \in \mathfrak{A}$ together with a morphism $c : X \to \mathcal{F}(X)$.

The object $X$ is often called state space and $c$ is called the transition or coalgebra structure. $X$ represents the kind of data that characterize the system, $\mathcal{F}(X)$ represents the interface by means of which we can interact and we can see the system and $c(\cdot)$ represents the dynamics of the system and it links the evolution of the state to the I/O interface. Intuitively, coalgebras describe general state-based systems provided with dynamics given by $c$. The codomain of $c(\cdot)$ is often called the interface of the coalgebra.

The coalgebraic framework of [1] aims to describe a software program in terms of generated (I/O) behavior instead of data processing or symbolic manipulation. In particular, [1] describes the o-o concepts of class specifications and class implementations within the coalgebraic framework. Class specifications are linguistic entities consisting of three parts describing: (1) the methods; (2) the logical assertions that the methods have to satisfy; (3) the conditions that should hold for newly created objects. Class specifications define the features of a class in general without any reference to the specific data type which constitutes the state space of the class and without any specific interpretation of the methods of the class. Therefore, a class specification defines what behavior a class should be able to reproduce and NOT how it has to be designed to do it. Class implementations are the interpretation of class specifications and, basically, correspond to class definitions in o-o programming languages. Thus, class implementations define the particular type of data that constitute the state space and the methods that implement those given by the class specification. Formally, a class implementation is given by a triple $(X, c : X \to T(X), x_0)$, where $X$ is an interpretation of the state space $X$ of the class specification, $c : X \to T(X)$ is a coalgebra that gives an interpretation of the methods, that is it consists of a set of functions that implement the methods defined in the specification. Finally $x_0 \in X$ is an element which should satisfy the creation conditions reported in the class specification.

We can access to the states only through the methods; thus, once a class implementation is given, it is important to identify the states which are indistinguishable by the methods. This is formalized in [1] by giving a specific definition of bisimulation, an equivalence relation on state spaces which implies the equivalence of the external behavior of two classes. Given a class specification $S$ it is possible to define a category $\text{Class}(S)$ of implementations of the specification. Summarizing the formal details of [1], the objects of $\text{Class}(S)$ are all the implementations of the specification $S$ and the morphisms link implementations whose external behavior is indistinguishable using the methods (i.e. they preserve bisimulation). Within this coalgebraic modeling framework it is also possible to define the concept of inheritance [1], both between class specifications and between class implementations. A class
specification S inherits from a class implementation T if all the methods, assertions and creation conditions of T form part of S, but S may be extended with (1) additional methods, (2) additional assertions (or strengthened assertions of T), (3) additional creation conditions (or strengthened creation conditions of T), (4) extending the output type A of an attribute method in T to a super-type A′ ⊇ A and the input type B of a procedure method in T to a super-type B′ ⊇ B. These four points ensure that implementations of the child specification S are also implementations of the parent specification T, namely that the external behavior generated by T is contained in the external behavior generated by S. Formally, we say that S inherits T if and only if there is a forgetful functor F between the categories Class(S) and Class(T), whose role is to erase the extra structure we added in the specification S. Thus, a specification S inherits from a specification T if erasing some methods, deleting (or weakening) some assertions and creation conditions from any of its implementations we get an implementation of the specification T. It is also possible to define inheritance also for class implementations: a class implementation B ∈ Class(S) inherits from a class implementation A ∈ Class(T) if there is a morphism of classes f : F(B) →→ A in the category Class(T). We shall call B a subclass of A. For further details see [1].

3. A categorical view of dynamical systems

Since within the coalgebraic framework for software classes the focus is on the generated behavior rather than on the functional implementation, we believe that a key ingredient for the extension of this modeling language to dynamical systems is the concept of interface behavior as in the behavioral approach. The basic idea in our coalgebraic description of dynamical systems is that a dynamical system can be seen as a class that implements a certain interface behavior. Recalling the previous definitions, a dynamical system specification defines the interface behavior that a dynamical system must be able to reproduce and the ways in which it is possible to interact with the system (the methods), while a dynamical system implementation represents a particular way to generate the specified behavior.

A dynamical system specification consists of three parts: methods, assertions and creation conditions. As for software classes, methods represent both a way through which we can get information about the system state and a way by which we can influence the system’s evolution. A fundamental difference between software classes and dynamical systems is the fact that dynamical systems is that they evolve synchronously with respect to an ordered time set (e.g. ℜ, ℤ, etc.), changing the information presented at the external interface at each time step (which can be infinitesimal). This also means that for dynamical systems the methods that determine their generable behavior have to be invoked at each instant.

A very large number of dynamical systems can be represented by considering only two kinds of methods: the attribute method and the procedure method. Let U and Y be a defined input and output space respectively and let X be the “unknown” state space and T be an unknown ordered time set. method, renamed evolve, are given by:

\[ \text{out} : X \mapsto Y \quad \text{evolve} : X \times U \times T \mapsto X \]

Thus, the out method maps the state into the output space and the method evolve encodes the evolution of the system and it depends on the time set T. The role of the input is to parameterize the evolution of the system. The method specification says that we can see the state space only in the output space Y, that the system has to be able to evolve with respect to a time set and that we can influence the evolution by elements in the set U. consist of a particular I/O behavior B that the system must be able to reproduce. This means that the methods of the dynamical system have to be related in such a way that it is possible to reproduce the behavior B; it is possible to select a particular I/O trajectory in the behavior by selecting a proper input trajectory. Finally, the creation conditions represent the initial output condition that have to be satisfied by the implementations of the specification at an instant that is chosen to be the initial one.

Dynamical systems implementations are the interpretation of dynamical systems specifications and they define the particular structure of the system. They are defined by the triple \((X, T), c : X \times T \mapsto T(X), m_0\), where \(T(X)\) represents the I/O interface of the system. The choice of \(X\) and \(T\) implies the choice of a structure for the state space and for the ordered time set of the dynamical system. The coalgebra \(c(\cdot)\) consists of a set of maps related in such a way to allow to implement the behavior requested in the assertions compartment of the specification. Basically the coalgebra defines the output map, which is associated to the method \(\text{out}\), \(h : X \mapsto Y\) and the evolution map, which is associated to the method \(\text{evolve}\), \(\Phi : X \times U \times T \mapsto X\).

The choice of the output map and of the evolutionary vector field has to be such that for every possible pair \((u(\cdot), y(\cdot))\) we have that

\[ (u(t), y(t)) \in B \]

In other words we require that the system implements the specified behavior. Finally, \(x_0 \in X\) is the initial state of the implementation and it has to be such that the initial output configuration satisfy the creation conditions reported in the specification.
Using the proposed coalgebraic framework, it is possible to extend the notion of inheritance to dynamical systems. In Sec. 2 we reported a set of 4 possible extensions through which it is possible to build a child specification. We will now restate these possible extensions giving them a more system theoretic interpretation and defining the concept of inheritance for dynamical systems specification. A dynamical system specification $D$ inherits from a dynamical system specification $E$ if all the methods, assertions and creation conditions of $E$ form part of $D$, but:

1. $D$ may have additional methods. It is possible to add outputs methods and to add evolution methods. In the latter case we allow the state to evolve in several ways; only one evolution method must be invoked at a certain instant meaning that the state can evolve following only one integral curve of an evolutionary vector field at a time.
2. $D$ may have additional assertions and, moreover, the assertions of $E$ may be strengthened. For dynamical systems, if $B$ is the behavior required by the specification $E$, it means that the child system could implement a behavior $B_1 \supseteq B$. In this way all the trajectories of the parent behavior can be implemented by the child and, therefore, the LSP is satisfied.
3. $D$ may have additional creation conditions and, moreover, the creation conditions of $E$ may be strengthened. This extension allows to the derived system to impose further conditions on the value of the initial outputs.
4. The output space $Y$ and the input space $U$ may be extended to super-spaces $Y' \supseteq Y$ and $U' \supseteq U$ respectively. It is therefore possible to extend the output space to support more kind of attributes and the input space to provide a further capability to parameterize the evolution of the state.

These four points imply that implementations of the child specification are also implementations of the parent specification and that children systems can be used in place of parent system without any loss of performance, but with an extra flexibility and extra functionalities. This property can be formalized in terms of category theory by formulating the existence of a forgetful functor between the categories $\text{Class}(S)$ and $\text{Class}(T)$ as described in Sec. 2 for software classes. It is possible to straightforwardly apply the construction provided for classes implementations to dynamic systems implementations. Thus we have that an implementation $B$ of the child specification $D$ is a subclass of an implementation $A$ of the parent specification $E$ if, after having forgotten all the extensions introduced in the inheritance process, the external behavior of the class $\mathcal{F}(B)$, where $\mathcal{F}$ is the forgetful functor, is the same as that of the system $A$.

4. Conclusion

We have presented an o-o modeling framework for dynamical systems based on category theory and on the behavioral approach. This framework allows to formally define the concept of inheritance between classes of dynamical systems, with a strong emphasis on behavioral conformity. We consider the latter property very important in order to introduce the principle of design by extension in the domain of mechatronic system. Our future work aim to analyze significant cases of study, in which the proposed extension mechanism will highlight its usefulness for the design of “enhanced” mechatronic components. Loosely speaking, a complex manufacturing system can be thought as an interconnection of basic components, either logical or physical, which exchange information [5]. The overall behavior of the system is determined by the fact that each component is able to reproduce at its I/O port a given behavior. Given a component of a complex system, either logical or physical, exploiting the inheritance framework proposed in the paper, it is possible to build a derived component that is able to implement some extra functionalities with respect to the parent component but that, at the same time, it is able to exactly reproduce the I/O behavior of the parent component and, consequently, to allow the overall manufacturing system to work as with the parent component.

References

A General SPICE Model of the Stochastic Resonance Process

Gabriel Vasilescu† and Luonan Chen‡

† LISIF, University Pierre & Marie Curie (Paris VI) 3 rue Galilée, 94200 Ivry/Seine, France
‡ Dept. of Electrical Eng. and Electronics, Osaka Sangyo University Nakagaito 3-1-1, Daito, Osaka, Japan 574-8530 Email: gv@lisif.jussieu.fr, chen@elec.osaka-sandai.ac.jp

Abstract – In the last years, it has been seen a growth of interest in Stochastic Resonance (SR) phenomenon, which becomes popular in various fields. This contribution presents a PSPICE macro-model of the stochastic resonance process, by means of which we can determine the intensity of the added noise and its characteristics, in order to enable the detection of under-threshold signals (stimuli).

1. Background

Stochastic resonance (SR) is a phenomenon observed in both natural and man-made non-linear systems, requiring the superposition of some external noise to the useful signal. Contrary to linear systems, where noise deteriorates the performance, in non-linear systems detectability can be improved by properly tuning the intensity of the additive external noise. A frequency domain approach shows that the external noise and the periodic signal interact, pumping power from the whole noise spectrum into a single mode, reinforcing the useful signal. To occur, the SR process needs four ingredients:

- A non-linear system, with a threshold operation and two stable states.
- A sub-threshold useful signal (stimulus), often periodic.
- An external random noise source, which superposes onto the useful signal.
- A performance criterion, to evaluate the signal delivered at the output of the non-linear system.

In the time domain, the mechanism of SR can be explained as follows: in the absence of noise, the amplitude of the stimulus never crosses the threshold of the non-linear system, whose output does not react. By adding random noise to the useful signal, at least in some time intervals its amplitude increases enough to overpass the threshold and hence it triggers the non-linear system. Consequently, the output of the non-linear system switches between its two stable states and delivers a noisy signal, from which further processing can recover most of the information carried by the stimulus. In this way, adding an appropriate amount of noise (with a known statistics) is beneficial to detecting a weak signal. Note that adding too much noise degrades the operation, which fails to extract the useful information.

This phenomenon has been observed and reported in a variety of fields covering biological sensing systems, chemistry, genetics, neuroscience, etc. For instance, it appears in feeding or predator avoidance of some animal species (such as fish), in auditory and vision systems operation (when perception threshold is impaired by age), in human tactile sensation improvement by adding random vibration [1], etc.

Later, this concept has found application in electronic engineering, namely in signal processing (to extract a useful signal embedded in noise) and in system control. Equally, SR has been used to explain periodic recurrences in the Earth’s ice ages. Recently, this phenomenon is suspected to appear in the mechanism of cellular communications [4], where it may play an important role. Several fundamental experiments have shown that intercellular and intracellular communications are affected either by the noise arising inside the cell (due to fluctuations in birth and dead rate of individual molecules), or external noise (due to environment fluctuations). In contrast to the former, the latter, being common to all cells, can facilitate communication between two populations of cells and therefore induce a cooperative behaviour, by imposing some synchronization.

Being so important in various fields where laboratory experiments are very difficult, very expensive or even impossible to achieve, it is paramount to find an electrical model of the stochastic resonance process, by means of which simulation can be easily carried out with a commonly used program (like SPICE). Provided that an analogy between the electrical domain and the field of the particular application can be established, simulation is a valuable tool not only to avoid high – cost experiments, but equally to give a deeper insight into the mechanism of stochastic resonance process.
2. Model Structure

The functional diagram of the proposed macro-model is presented in Fig. 1. According to the explanation proposed in the time-domain, the model must include the following elements:

- An operational amplifier is used as comparator. Its threshold value is set by means of an adjustable DC reference source connected to one of its inputs.
- The useful signal (stimulus) is delivered by a sinusoidal generator.
- A Colpitts-like chaotic oscillator [2], whose behaviour is similar to a Rössler attractor, delivers the noise to be added (a chaotic signal, with a strong periodic component).
- In order to avoid mutual influence when both signals are added, a buffer follows the noise generator, as well as the stimulus generator.
- Both buffers employ LM324 operational amplifiers, connected as unity gain stages (followers).
- Concerning the adder, the simplest circuit with two resistors has been selected. By adjusting the values of these resistors one can modify the weighting factors of the two terms in the sum, and so the amount of added noise.

The detailed schematic of the proposed SR macro-model is presented in Fig. 2. Note the following points:

- The frequency of the sinusoidal generator is set to 1 kHz and its amplitude to 2V.
- The threshold voltage has been set to 2.1V. Consequently, in the absence of noise, the sinusoidal generator alone is unable to switch the output of the comparator.
- The chaotic oscillator delivers a signal with random amplitude, ranging between 0 and 4.5V. Its frequency is around 38 kHz.
Fig. 3 Typical waveforms obtained with PSPICE.

The performance criterion is established relatively to the output signal, which must simultaneously fulfill two conditions:

- It must be periodic, with the same frequency as that of the stimuli (sinusoidal generator).
- Its amplitude must be sufficiently high (i.e., the comparator must switch between its two stable states, delivering an output with a high amplitude).

3. Results

A PSPICE simulation, with the values of components proposed in Fig. 2, has been performed. The general form of various waveforms is presented in Fig. 3. V(STIM) refers to the waveform of the sinusoidal generator (at its buffer output). V(NOISE) is the waveform delivered by the chaotic oscillator (at its buffer output). V(OUT) is the voltage waveform at the comparator output (note that it switches between -10 V and +10 V).

By comparing V(OUT) with V(STIM), it is obvious that the former has the same frequency as the stimulus, and consequently the first condition imposed by the performance criterion is fulfilled.

The second condition (output signal with large amplitude) is equally satisfied, since the supply voltages (V+ and V-) set the output swing.

4. Discussion

Although fairly different from a sinusoid, V(OUT) is a large amplitude signal, which hopefully enables the neuronal processing and retrieving of the stimulus frequency.

In order to understand what happens in a biological system, let us consider the case of a fish whose survival depends on its ability to early detect the predator approach.

The fish, by means of its external sensor areas, can sense the vibrations present in its environment, then this information is transmitted to the neurons specialized to process it. Very likely, the fish must be able to identify from the gathered information the predator signature (early, when it is still out of range) and to react rapidly.

Evidence exists that the paddlefish has a peripheral sensory system that acts in this way and takes advantage of the stochastic resonance. Experiments on its behaviour have been conducted [3], and they show that adding external noise enhances the perception of “friend or foe” approach.

The following question arises now: what is the appropriate amount of noise to be added, in order to enable stimuli perception by stochastic resonance. Simulation represents a very elegant and non-expensive tool to answer. Transposed into the electrical domain, this means to investigate the effect of modifying the signal-to-noise ratio (SNR) at the comparator input. The simplest way to achieve this is to adjust the values of resistors RN, RS and to check each time whether the output waveform satisfies the performance criterion or not.

We adopt as reference the case when RN = RS = 1 kΩ (equal weighting factors, the resulting signal to
noise ratio being denoted by SNR0). The corresponding waveforms are shown in Fig. 4 and they are similar to those presented in Fig. 3, excepting for the time scale, which is different. As the output signal has comfortable amplitude (the swing is between –10V and about +9V), and it has the same frequency, the performance criterion is fulfilled. This represents a good choice.

The waveforms given in Fig. 5 correspond to the case when the amount of added noise is kept constant, but the stimuli signal has been attenuated with respect to noise (RN = 1 kΩ and RS = 5 kΩ). In other words, this time noise dominates the signal and consequently the SNR is deteriorated. It is obvious that the output signal does no more switch between the two stable states (actually, it is comprised between +3V and +9V). This may lead to impossibility of identifying stimuli signature by the biological system. Furthermore, since the output waveform is now very noisy, the frequency of the stimuli is rather difficult to be recovered.

Fig. 6 Case when intensity of additive noise is too weak.

A quite different situation is depicted in Fig. 6, where the waveforms refer to the case when the amount of added noise is so weak, that stochastic resonance is only incipient (RN = 5 kΩ and RS = 1kΩ). Here the amplitude of the input signal is quite small, it never triggers the comparator; therefore, no output swing from negative to positive voltages occurs. This may seriously impair the ability of further processing and extracting stimuli signature.

Of course, if the added noise is still more reduced, stochastic resonance fairly disappears. For instance, for RN = 20 kΩ (or greater) and RS = 1 kΩ, no pulses exist at the comparator output, which remains flat at –10 V.

We have checked the robustness of the proposed model by replacing the sinusoidal stimulus with the signal delivered by a pulse generator. No particular incidence has been noted, provided that the stimulus frequency is chosen much smaller than the frequency of noise. Finally, a Vilnius chaotic oscillator [5] has been substituted to its Colpitts-like counter-part; a sample of the obtained waveforms is presented in Fig. 7.

5. Conclusion

A general, versatile, robust SPICE model of the stochastic resonance has been proposed. This model is called general, since it is not dependent on a particular case study or on a particular application field. It may be used as a tool to simulate the stochastic resonance appearing in physics, in a chemical reaction, or in a biological process, provided that a one-to-one correspondence is previously established between the components of the SPICE model and the elements of the system under investigation.

This model is expected to give deeper insight into the dynamics of the investigated system, by avoiding to repeat expensive, time-consuming or difficult experiments.

References


A new approach to predict the possible bit streams of a Sigma Delta modulator.

Emer Condon † Paul Curran † and Orla Feely‡

†School of Electronic, Electrical and Mechanical Engineering, University College Dublin, Belfield, Dublin 4, Ireland
Email:emer.condon@ee.ucd.ie, paul.curran@ucd.ie and orla.feely@ucd.ie

Abstract—The possible output bit streams of a first order Sigma Delta modulator with integrator leakage, for any periodic input is investigated. Given certain periodic inputs, it is shown that not all of the possible output sequences can occur. In particular, it is proven that if the input frequency is rational i.e. $\omega = \frac{\pi}{q}$, that only bitstreams which are dependent on this frequency are possible.

1. Introduction

$\Sigma\Delta$ modulators are widely used in analog-to-digital and digital-to-analog conversion. Applications range from cellular phone technology to ultra sound imaging and more recently are used in Super Audio CD (SA-CD) [1]. However, despite their widespread use a complete understanding of their operation does not exist. This is due to the presence of the nonlinear element- the quantiser [2]. Essentially, there has been two approaches to analysing Sigma Delta modulators. The first is to linearise the nonlinearity and apply standard linear theory [2]. The second is to use the theory of nonlinear dynamics. However, the majority of this work concentrates on the case of a DC input [3]. In this paper, a new approach to investigating Sigma Delta modulator with a periodic input is presented. In contrast to work previously done which examined the behaviour of the state-space variables [4], this method looks at the output bit stream of the system. In particular, it considers successive blocks of positive ones and negative ones and determines constraints on the size of these blocks. Viewing the system in this way enables us to prove results regarding the form of the output sequences for given inputs. Information regarding the output sequences can help indicate the presence of limit cycles [5]. In addition, a greater knowledge regarding the output sequences can help indicate the presence of limit cycles [5]. In addition, a greater knowledge of the output sequences can help in the design of more efficient decoders [6]. In this paper, we focus on the first order Sigma Delta modulator. We begin by introducing the concept of $m$ sequences where $m$ and four lemma which are necessary to determine constraints on the $m$ sequence. In section 3, the Sigma Delta modulator under certain conditions is examined. Results regarding the $m$ sequences are presented. These are representative of the type of results which can be obtained by using this new approach. While this paper considers only the first order Sigma Delta modulator with integrator leakage, it is hoped that it will provide a frame work for the investigation of higher order systems.

2. The Sigma Delta modulator with integrator leakage

2.1. The first order Sigma Delta modulator

![Figure 1: First Order Sigma Delta modulator with integrator leakage](image)

The first order $\Sigma\Delta$ modulator is shown in Fig.1. The equation describing the $\Sigma\Delta$ modulator with a periodic input, is given as follows:

$$
\theta_n = \theta_0 + n\omega \mod(2\pi)
$$

$$
u_{n+1} = pu_n + f(\theta_n) - \text{sign}(u_n)
$$

where

$$
\text{sign}(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0 \end{cases}
$$

While any periodic input could be considered, in this paper the input to the modulator is $f(\theta_n) = \alpha \sin(\theta_n)$ with $|\alpha| < 1$, $\theta_n \in [0, 2\pi)$. The effect of integrator leakage is incorporated into the system by taking $p < 1$. In general, $\Sigma\Delta$ modulators are designed such that the pole of the system is as close to 1 as possible. Hence, in this paper both values of $p$ which are equal to 1 and just slightly less than 1, e.g. $p = 0.99$ are considered.

The output of the system is defined as:

$$
y_n = \text{sign}(u_n)
$$

2.2. The critical upper and lower bounds

Consider the set $\Gamma$ defined as

$$
\Gamma = \left\{ \left[ \begin{array}{c} u \\ \theta \end{array} \right] \in \mathbb{R} \times [0, 2\pi] : \frac{-1}{p} + f(\theta) \leq u \leq \frac{1}{p} + f(\theta) \right\}
$$
From [4], it is shown that the set $\Gamma$ is invariant and globally attracting. Assume that $u_0 \in \Gamma$ and $u_0 \geq 0$. Consider the following output of the system as defined in (2) $\rho \ni 1 - 1 - 1 - 1 - \cdots$ where $m_i$ is defined as the number of positive or negative ones in the output before a switch, e.g. $m_0 = 4$ in the above sequence. This sequence of $m_i$ is referred to as the $m$-sequence and for simplicity will be considered throughout the paper rather than $y_n$. It is assumed that $y_0 = 1$. It follows that $y_n = 1$ for $0 \leq n < m_0$ and $y_n = -1$ for $m_0 \leq n < m_0 + m_1$. Define:

$$\rho_n = \sin(\theta_0 + n\omega) \quad \text{for} \quad n \geq 0$$  \hspace{1cm} (3)

$$B_0 = 0, B_{n+1} = \frac{1}{p^n} + B_n, \quad \text{for} \quad n \geq 0$$  \hspace{1cm} (4)

$$C_0 = 0, C_{n+1} = \frac{\rho_n}{p^n} + C_n, \quad \text{for} \quad n \geq 0$$  \hspace{1cm} (5)

In this section, we present four lemma regarding the bounds on $u_0$. Lemma 1 states the appropriate equations for $u_0$ in the ranges $0 \leq n \leq m_0$ and $m_0 \leq n \leq m_0 + m_1$. From this, we can determine the bounds on $u_0$. Lemma 2 proceeds to show that only one of the bounds on $u_0$ in each range is critical. The proofs of these lemma are straightforward and are omitted due to restrictions on space.

**Lemma 1**

For system (1) with $u_0 \geq 0$

$$u_n = p^\rho u_0 + \alpha p^\rho C_n - p^\rho B_n \quad \text{for} \quad 0 \leq n \leq m_0$$

$$u_n = p^\rho u_0 - p^\rho m_0 B_n + \alpha p^\rho C_n + p^\rho m_0 B_{n-m_0} \quad \text{for} \quad m_0 \leq n \leq m_0 + m_1$$

It follows that

$$u_0 \geq B_n - \alpha C_n \quad \text{for} \quad 1 \leq n \leq m_0 - 1$$

and

$$u_0 < B_{m_0} - \alpha C_n - \frac{1}{p^n} B_{n-m_0} \quad \text{for} \quad m_0 \leq n \leq m_0 + m_1 - 1$$

In Lemma 2, we find that in a given range there exists a value of $n$ which yields a critical bound. This is true provided $|\alpha| < 1$

**Lemma 2**

If $|\alpha| < 1$ then

$$u_0 \geq B_n - \alpha C_n \quad \text{for} \quad 1 \leq n \leq m_0 - 1$$

iff

$$u_0 \geq B_{m_0-1} - \alpha C_{m_0-1}$$  \hspace{1cm} (6)

and

$$u_0 < B_{m_0} - \alpha C_n - \frac{1}{p^n} B_{n-m_0} \quad \text{for} \quad m_0 \leq n \leq m_0 + m_1 - 1$$

iff

$$u_0 < B_{m_0} - \alpha C_{m_0+m_1-1} - p^\rho m_0 B_{m_1-1}$$  \hspace{1cm} (7)

We now show that these lemma are true for the more general case where $s \geq 1$.

**Lemma 3**

For system (1) with $u_0 \geq 0$

$$u_n = p^\rho u_0 + \alpha p^\rho C_n - p^\rho B_n + p^\rho m_0 B_{m_1} + \cdots$$

$$\cdots = p^\rho m_0 + \cdots + m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})}$$

for

$$m_0 + \cdots + m_{2s-1} + 1 \leq n \leq m_0 + \cdots + m_{2s}$$

with $s \geq 1$ and

$$u_n = p^\rho u_0 + \alpha p^\rho C_n - p^\rho B_n + p^\rho m_0 B_{m_1} + \cdots$$

$$\cdots = p^\rho m_0 + \cdots + m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})}$$

for

$$m_0 + \cdots + m_{2s-2} + 1 \leq n \leq m_0 + m_1 + m_2 + \cdots + m_{2s-1}$$

with $s \geq 1$.

It follows that

$$u_0 \geq B_{m_0} - p^\rho m_0 B_{m_1} + \cdots + p^\rho m_0 \cdots m_{2s-1} B_{m_{2s-1}}$$

$$+ p^\rho m_0 \cdots m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})} - \alpha C_n$$

for

$$m_0 + \cdots + m_{2s-1} + 1 \leq n \leq m_0 + \cdots + m_{2s} - 1$$

and

$$u_0 < B_{m_0} - p^\rho m_0 B_{m_1} + \cdots + p^\rho m_0 \cdots m_{2s-1} B_{m_{2s-1}}$$

$$- p^\rho m_0 \cdots m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})} - \alpha C_n$$

for

$$m_0 + \cdots + m_{2s-2} \leq n \leq m_0 + \cdots + m_{2s-1} - 1$$

Again, as in Lemma 2, we now find the critical bound in each range.

**Lemma 4**

If $|\alpha| < 1$ then

$$u_0 \geq B_{m_0} - p^\rho m_0 B_{m_1} + \cdots$$

$$+ p^\rho m_0 \cdots m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})} - \alpha C_n$$

for

$$m_0 + \cdots + m_{2s-1} + 1 \leq n \leq m_0 + \cdots + m_{2s} - 1$$

iff

$$u_0 \geq B_{m_0} - p^\rho m_0 B_{m_1} + \cdots + p^\rho m_0 \cdots m_{2s-1} B_{m_{2s-1}}$$

$$+ p^\rho m_0 \cdots m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})} - \alpha C_n$$

for $s \geq 1$ and

$$u_0 < B_{m_0} - p^\rho m_0 B_{m_1} + \cdots + p^\rho m_0 \cdots m_{2s-1} B_{m_{2s-1}}$$

$$- p^\rho m_0 \cdots m_{2s-1} B_{N-(m_0 + \cdots + m_{2s-1})} - \alpha C_n$$
for \( m_0 + \cdots + m_{2r-2} \leq n \leq m_0 + \cdots + m_{2r-1} - 1 \)

iff \( u_0 < B_{m_0} - p^{-m_0}B_{m_1} + \cdots - p^{-m_0-\cdots-m_{2r-2}}B_{m_{2r-1}} - 1 \)

\[ -\alpha C_{m_0+\cdots+m_{2r-1}} \] (9)

for \( s \geq 1 \)

Hence, the critical bounds on \( u_0 \) have been determined. In the case where \( p = 1 \) in (1), then (8) and (9) take the simpler form as follows:

\[ u_0 \geq (m_0 - \cdots - m_{2r-1}) - \alpha p_{m_0} + \cdots + p_{m_{2r-2}} \] (10)

and

\[ u_0 < (m_0 - \cdots - m_{2r-1}) - \alpha p_{m_0} + \cdots + p_{m_{2r-2}} \] (11)

In the next section, we will compare the critical bounds for different values of \( s \). This allows us to determine conditions on the \( m \) sequence.

3. Comparison of Bounds

Consider (1) where the input frequency is \( \omega = \frac{r\pi}{q} \), \( q \geq 2 \), \( 0 < r \leq q \) where \( r, q \in \mathbb{Z} \) are coprime and \( r \) is odd.

\[ \rho_{n+2q} = \rho_n \] \[ \rho_{n+q} = -\rho_n \] for all \( n \) (12)

By comparing the bounds on \( u_0 \) we find that there are restrictions on the size of \( m_i \) for \( i \geq 1 \). In fact, the size of \( m_i \) depends on the frequency of the input.

**Lemma 5**

For system in (1) with \( p = 1 \) and \( p \) close to 1 then \( m_i \leq q + 1 \) for all \( i \geq 1 \). In particular, for \( p = 1 \) \( m_i \leq q \).

**Proof:**
The proof is given in the appendix

**Lemma 6**

For system (1) with \( p = 1 \), if \( m_i = q \) for any \( i \geq 1 \) then \( m_j = q \) for all \( j \geq i \).

**Lemma 7**

For system (1) with \( p = 1 \), if \( m_i = m \) for any \( i \geq 1 \) then \( m \) must be a factor of \( q \).

**Example** Consider system (1) when \( p = 1 \) with the following parameters \( \theta_0 = 0 \), \( \omega = \frac{pi}{3} \) and \( \alpha = 0.8 \) and \( \alpha = 0.25 \). From simulations, we find that the possible \( m \) sequences as shown in Table 1 and Table 2. We require that the system is in steady state operation and so all \( u_0 \in \Gamma \). Therefore, we consider thousands of initial conditions which are in the range \([0,1.8]\) and \([0,1.25]\) respectively.

<table>
<thead>
<tr>
<th>( m_1 )</th>
<th>( m_2 )</th>
<th>( m_3 )</th>
<th>( m_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: M sequences for \( \omega = \frac{pi}{3}, \alpha = 0.8 \)

<table>
<thead>
<tr>
<th>( m_1 )</th>
<th>( m_2 )</th>
<th>( m_3 )</th>
<th>( m_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 2: M sequences for \( \omega = \frac{pi}{3}, \alpha = 0.25 \)

Discarding \( m_0 \), it is seen that the \( m \) sequences repeat and so in this case \( m_1 = m_2 = m_3 = \cdots \). Hence, \( m_1, \cdots, m_4 \) need only be considered here.

It is clear that the lemma are consistent with the results of the simulations. **Lemma 5** states that only \( m_i = 1, 2, 3 \) can occur for any \( i \geq 1 \). This is evident in Table 1 and 2. **Lemma 6** states that when \( m_i = 3 \) then all the following \( m_j = 3 \) for \( j \geq 1 \). This is evident in the third row in Table 1. **Lemma 7** states that if the \( m \) sequence is of the form \( m_i = m \) for any \( i \geq j \) then \( m \) must be a factor of \( q \). As can be seen, \( m_i = 2 \) for any \( i \geq 1 \) is not present. For smaller values of \( \alpha \), then a \( m \) sequence of the form \( m_i = 1 \) for any \( i \geq 1 \) can be observed in Table 2.

**Conclusions**

The paper discusses the allowed sequences at the output of the first order Sigma Delta modulator system with integrator leakage for particular inputs. We are currently in the process of extending this approach to higher order Sigma Delta modulator systems.

**Acknowledgments**

This research is supported by Science Foundation Ireland.

**References**


Appendix

Proof of Lemma 5
Comparing (8) and (9) for the same $s$ yields

$$B_{m_{2s-1}} < 1 + \alpha p^{m_{2s-1}}(C_{m_{2s-1}} - C_{m_{2s-1} - 1})$$

From (5)

$$p^{m_{2s-1}}(C_{m_{2s-1}} - C_{m_{2s-1} - 1}) =$$

$$p^{-m_{2s-1}}(\rho_{m0} + \ldots + m_{2s-2} + p\rho_{m0} + \ldots + m_{2s-3} + \ldots + p^{-m_{2s-1}}\rho_{m0} + \ldots + m_{2s} - m_{2s} + 1)$$

Let

$$\delta = (\rho_{m0} + \ldots + m_{2s-2} + p\rho_{m0} + \ldots + m_{2s-3} + \ldots + p^{-m_{2s-1}}\rho_{m0} + \ldots + m_{2s} - m_{2s} + 1)$$

The possible $\rho$ values are as follows:

$$\rho_{0}, \rho_{1}, \ldots, \rho_{q-1}, -\rho_{0}, -\rho_{1}, \ldots, -\rho_{q-1}$$

$$\delta = 0, 1 - p^{q} + p^{2q} - \ldots$$

But $\delta$ is a linear combination of $\rho_i$ for $0 \geq i \geq q - 1$ where all the coefficients have magnitude less than 1. This is clear by noting that

$$0 \leq |\delta| \leq q as |p| \leq 1 for all i. Therefore,$$

$$B_{m_{2s-1}} < 1 + \frac{q}{p^{m_{2s-1}}}$$

From (4)

$$\frac{1 - p^{m_{2s-1}}}{1 - p} < p^{m_{2s-1}} + q$$

which in turn can be written as:

$$m_{2s} < 1 + \frac{\ln \left( \frac{(2-p)}{1-q(1-p)} \right)}{\ln(1/p)}$$

Since $p$ is sufficiently close to 1, this becomes:

$$m_{2s} < 2 + q$$

Similarly,

$$m_{2s+1} < 2 + q$$

Therefore,

$$m_i < 2 + q$$

for $i \geq 1$.

In the case where $p = 1$, the bounds are given as in (10) and (11)

The lower bound and upper bounds can be written as follows:

$$u_0 \geq (m_0 - \ldots + m_2 s - m_{2s+1} + 1) - \alpha \sigma_{2s}$$

and

$$u_0 < (m_0 - \ldots + m_2 s - m_{2s+1} + 1) - \alpha \sigma_{2s+1}$$

where

$$\sigma = \left\{ \begin{array}{ll}
\rho_0 & \text{if } m_0 + \ldots + m_i - 2 \equiv 0 \mod (2q) \\
\rho_0 + \rho_1 & \text{if } m_0 + \ldots + m_i - 2 \equiv 1 \mod (2q) \\
\vdots & \\
\rho_{q-1} - 1 & \text{if } m_0 + \ldots + m_i - 2 \equiv 0 \mod (q-1) \\
0 & \text{if } m_0 + \ldots + m_i - 2 \equiv 2q - 1 \mod (q-1) 
\end{array} \right.$$  \hspace{0.5cm} (15)

Comparing the bounds for the same $s$

$$(m_0 - \ldots + m_2 s + m_{2s+1} + 1) - \alpha \sigma_{2s+1}$$

This yields that

$$m_{2s+1} < 2 - \alpha (\sigma_{2s+1} - \sigma_{2s})$$

Using (15) \(|\sigma_{2s+1} - \sigma_{2s}| \leq 2q$.

Hence, \(m_{2s+1} < 2 + 2q$$

If $m_{2s+1} = 1 + 2q$ then 17 is

$$1 + 2q < 2 - \alpha (\sigma_{2s+1} - \sigma_{2s})$$

Now if $m_{2s+1} = 1 + 2q$ then $m_{2s+1} \equiv 1 \mod (2q)$, the congruence of $\sigma_{2s+1}$ and $\sigma_{2s}$ differ by one.

Hence \(|\sigma_{2s+1} - \sigma_{2s}| \leq 1 so 17 becomes $1 + 2q < 3 i.e. q < 1$ Therefore, $m_{2s+1} \neq 1 + 2q$

More generally, if $m_{2s+1} = 2q - k for k \geq 0$ then it follows that $2q - k < 2 - \alpha (\sigma_{2s+1} - \sigma_{2s})$

But $m_{2s+1} = -k \mod (2q)$ which means that \(|\sigma_{2s+1} - \sigma_{2s}| \leq k for any k \leq q\)

Therefore, $2q - k < 2 + k i.e. q < 1 + k$

This fails for $k \leq q - 1$. Therefore, $m_{2s+1} \neq 2q - k for 0 \leq k \leq q - 1$

Therefore, $m_{2s+1} \leq q$. Similarly, $m_{2s} \leq q.$
Two Types of Fuzzy $c$-Means for Data with Tolerance

Yuchi Kanzawa†, Yasunori Endo‡, Sadaaki Miyamoto‡
†Department of Communication Engineering,
Shibaura Institute of Technology
Email: kanz@kanz.ce.shibaura-it.ac.jp
‡Department of Risk Engineering,
Faculty of System and Information Engineering,
University of Tsukuba

Abstract—In this paper, a new clustering algorithm is proposed for data with some errors. It is based on the entropy based fuzzy $c$-means. The first, the tolerance which means the range of the error is introduced into optimization problems related with clustering in 2 different types of way. One is that the tolerance minimizes the corresponding objective function, which has already been proposed, and the other is that the tolerance maximizes one. These algorithms are constructed based on the results of solving such optimization problems with Kuhn-Tucker conditions. Through some numerical experiments, robustness of clustering is considered.

1. Introduction

Fuzzy $c$-means (FCM) is very famous and representative method in clustering algorithms [1]. The FCM is based on hard $c$-means (HCM) and has been constructed by fuzzification of HCM. Some FCMs is used in the field of clustering. Each FCM corresponds with the way to fuzzy the HCM. Particularly, the entropy regularized FCM [2] is known as effective in FCMs.

There are many cases that data has some errors in clustering. The errors have been represented by inter-val values [3, 4] in these case. On the other hand, some of authors have formulated these error problems into the optimization problems with inequality constraints and have constructed new clustering algorithms through solving the problems [5]. In these algorithms, the variable for error minimizes the corresponding objective function.

In this paper, we propose new algorithms in the way that the variable for error maximize the corresponding objective function. The first, we will define the tolerance $\varepsilon_k$ which means the range of the error and formulate optimization problem within the tolerance. The next, we will construct new algorithms by solving the problems by using Kuhn-Tucker conditions. By some numerical examples, we conclude that applying both the former algorithm [5] and the algorithm proposed in this paper to the same data guarantees the robustness of the result of clustering.

2. Preliminaries

In this section, some notations and definitions are introduced. In the first subsection, the data for clustering and the membership by which each datum belong to each cluster are defined. In the second subsection, entropy based FCM for data with tolerance is introduced. This algorithm is the basis of our main theme.

2.1. Notations

In this subsection, the data for clustering and the membership by which each datum belong to each cluster are defined. The data set $x = \{x_1, x_2, \ldots, x_N\}$ is given. The membership by which $x_i$ belongs to cluster $j$ is denoted by $u_{i,j}$ ($i \in \{1, \ldots, N\}, j \in \{1, \ldots, C\}$). The constraint for $u$ is

$$\sum_{j=1}^{C} u_{i,j} = 1.$$ 

2.2. FCM for data with Tolerance

In this subsection, we introduce entropy-based FCM with tolerance [5] on which our result is based on. The cluster center set is denoted by $v$ with elements $v_j$ ($j \in \{1, \ldots, C\}$). The tolerance for the data $x$ is denoted by $\varepsilon$ with elements $\varepsilon_i$ ($i \in \{1, \ldots, N\}$). The maximum tolerance is denoted by $\kappa$ with non-negative elements $\kappa_i$ ($i \in \{1, \ldots, N\}$).

Entropy-based FCM with tolerance is for the following optimization problem:

$$\text{minimize}_{u, \varepsilon, v} J_{\lambda, \varepsilon} \quad \text{subject to} \quad \sum_{j=1}^{C} u_{i,j} = 1, \quad \|\varepsilon_i\|^2 \leq \kappa_i^2 \quad (\kappa_i > 0).$$
For this problem, entropy-based FCM with tolerance is described as the following algorithm.

Algorithm 1

1. **Give the number of cluster $C$, the maximum tolerance $\kappa_i$ s ($i \in \{1, \cdots, N\}$) for data $x_i$ s ($i \in \{1, \cdots, N\}$) and the value $\lambda$. Set the initial values of $\varepsilon$ and $v$.**

2. **Calculate $u$ such that**

   $$u_{i,j} = \left(\sum_{k=1}^{C} \exp(-\lambda d_{i,k})\right)^{-1} \exp(-\lambda d_{i,j}),$$

   **where**

   $$d_{i,j} = \|x_i + \varepsilon_i - v_j\|^2.$$

3. **Calculate $\varepsilon$ such that**

   $$\varepsilon_i = -\alpha_i \left( x_i - \sum_{j=1}^{C} u_{i,j} v_j \right),$$

   **where**

   $$\alpha_i = \min \left\{ \kappa_i \left\| x_i - \sum_{j=1}^{C} u_{i,j} v_j \right\|^{-1}, 1 \right\}.$$

4. **Calculate $v$ such that**

   $$v_i = U_i^{-1} \sum_{j=1}^{N} u_{j,i} (x_j + \varepsilon_j),$$

   **where**

   $$U_i = \sum_{j=1}^{N} u_{j,i}.$$

5. **Check the stopping criterion for ($u, \varepsilon, v$). If the criterion is not satisfied, go back to Step 2.**

3. Another Type of FCM for data with Tolerance

   In this section, we propose another type of entropy based FCM with tolerance in the way that the tolerance maximizes the objective function, which is considered the worst case with the tolerance. In the first subsection, the optimization problem is described and its KKT conditions are lead. In the second subsection, An iterative algorithm is proposed using these KKT conditions.

### 3.1. Optimization Problem and its KKT conditions

In this subsection, the optimization problem is described and its KKT conditions are lead.

For $x, u, \varepsilon, v$, we consider the following optimization problem:

$$\begin{align*}
\text{minimize}_{u, v, \varepsilon} & \quad J_{\lambda, e} \\
= & \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|x_i + \varepsilon_i - v_j\|^2 \\
+ & \lambda^{-1} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \log(u_{i,j}), \quad (1) \\
\text{maximize}_{\varepsilon} & \quad J_{\lambda, e}, \quad (2)
\end{align*}$$

under

$$\begin{align*}
\sum_{j=1}^{C} u_{i,j} = 1, \\
\|\varepsilon_i\|^2 \leq \kappa_i^2 \quad (\kappa_i > 0). \quad (3)
\end{align*}$$

The Lagrange function for (1) and (3) $L^{(1)}_{\lambda, k}$ is as below:

$$L^{(1)}_{\lambda, k}(x_i, \varepsilon_i, v_j) = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|x_i + \varepsilon_i - v_j\|^2$$

$$+ \lambda^{-1} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \log(u_{i,j})$$

$$+ \sum_{i=1}^{N} \left( \gamma_i \left( \sum_{j=1}^{C} u_{i,j} - 1 \right) \right),$$

where $\gamma = (\gamma_1, \ldots, \gamma_N)$ is KKT vectors. From KKT conditions, we obtain the following necessary conditions that the solution should satisfy:

$$u_{i,j} = \left(\sum_{k=1}^{C} \exp(-\lambda d_{i,k})\right)^{-1} \exp(-\lambda d_{i,j}),$$

$$v_i = U_i^{-1} \sum_{j=1}^{N} u_{j,i} (x_j + \varepsilon_j),$$

where

$$d_{i,j} = \|x_i + \varepsilon_i - v_j\|^2,$$

$$U_i = \sum_{j=1}^{N} u_{j,i}.$$

These conditions are the same as one of Algorithm 1. Since $J_{\lambda, e}$ is convex for $\varepsilon$, such $\varepsilon$ as maximizing $J_{\lambda, e}$ exists only in the boundary of the constraint (4). Thus, the optimization problem (2) and (4) is reformed into the following problem:

$$\begin{align*}
\text{maximize}_{\varepsilon, \lambda, e} & \quad J_{\lambda, e} \\
\text{under} & \quad \|\varepsilon_i\|^2 = \kappa_i^2 \quad (\kappa_i > 0). \quad (5)
\end{align*}$$
The Lagrange function for (5) and (6) \( L^{(2)}_{\lambda,k} \) is as below:
\[
L^{(2)}_{\lambda,k}(x_i, \varepsilon_i, v_j) = -\sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \|x_i + \varepsilon_i - v_j\|^2
+ \lambda \sum_{i=1}^{N} \sum_{j=1}^{C} u_{i,j} \log u_{i,j}
+ \sum_{i=1}^{N} \delta_i \left( \|E_i\|^2 - \kappa_i^2 \right),
\]
where \( \delta = (\delta_1, \ldots, \delta_N) \) is KKT vectors. From KKT conditions, we obtain the following necessary condition that the solution should satisfy:
\[
\varepsilon_i = \frac{\kappa_i \left( x_i - \sum_{j=1}^{C} u_{i,j} v_j \right)}{\|x_i - \sum_{j=1}^{C} u_{i,j} v_j\|}.
\]

### 3.2. Algorithm

From the discussion in the previous subsection, we obtain the following iterative algorithm.

**Algorithm 2**

1. **Step 1** Give the number of cluster \( C, \) the maximum tolerance \( \kappa_i \) \( (i \in \{1, \ldots, N\}) \) for data \( x, s \) \( (i \in \{1, \ldots, N\}) \) and the value \( \lambda \). Set the initial values of \( \varepsilon \) and \( v \).

2. **Step 2** Calculate \( u \) such that

   \[
u_{i,j} = \left( \sum_{k=1}^{C} \exp(-\lambda d_{i,k}) \right)^{-1} \exp(-\lambda d_{i,j}),\]

   where

   \[d_{i,j} = \|x_i + \varepsilon_i - v_j\|^2.\]

3. **Step 3** Calculate \( \varepsilon \) such that

   \[
   \varepsilon_i = \frac{\kappa_i \left( x_i - \sum_{j=1}^{C} u_{i,j} v_j \right)}{\|x_i - \sum_{j=1}^{C} u_{i,j} v_j\|}.
   \]

4. **Step 4** Calculate \( v \) such that

   \[
v_i = U^{-1}_i \sum_{j=1}^{N} u_{j,i} (x_j + \varepsilon_j),\]

   where

   \[U_i = \sum_{j=1}^{N} u_{j,i}.
   \]

5. **Step 5** Check the stopping criterion for \((u, \varepsilon, v)\). If the criterion is not satisfied, go back to Step 2.

### 4. Numerical Examples

In this section, we show the result obtained by applying Algorithm 1 and Algorithm 2 to Polaris data shown in Figure 1.

We gave \( C = 3 \) and \( \lambda = 0.1 \) and we set that \( \varepsilon_i = 0 \) \( (i \in \{1, \ldots, 51\}) \) and \( v_j \) \( (j \in \{1, \ldots, 3\}) \) are random from 0 to 100 as the initial values respectively. Figure 2 shows the result obtained by Algorithm 1 with \( \kappa_i = 1 \) \( (i \in \{1, \ldots, 51\}) \) and Figure 3 shows the result obtained by Algorithm 2 with \( \kappa_i = 1.0 \) \( (i \in \{1, \ldots, 51\}) \).

In these figures, 3 symbols +, x and * show kinds of clusters and bottoms/tops of arrows show \( x_i s/x_i + \varepsilon_i s \). We can see that both algorithms made the same clusters each other. While Algorithm 1 moved data far from their own cluster centers, Algorithm 2 moved data far from their own cluster centers.

Figure 1: Polaris Data

Figure 2: Algorithm 1 with \( \kappa_i = 1 \)
Algorithm 2 could not stop within 100000 iterations. We consider this is because $\kappa_i$'s are too large to make a certain classifying.

From the above numerical examples, we consider that obtaining the same result by both Algorithm 1 and Algorithm 2 with certain $\kappa_i$'s shows the stable result for these $\kappa_i$'s. Thus, we conclude that applying both algorithms to the same data guarantees some robustness for the errors of data.

5. Conclusion

In this paper, we proposed a new clustering algorithm. This algorithm are based on the entropy regularized FCM and can treat the data with some errors which is represented by $\varepsilon_i$'s. While $\varepsilon_i$'s minimize the corresponding objective function in the previously proposed algorithm [5], $\varepsilon_i$'s maximize the corresponding objective function in the algorithm proposed in this paper. Thorough numerical examples, we concluded that applying both algorithms to the same data guarantees some robustness for the errors of data. We add to this paper that the similar algorithm can be constructed for standard FCM with tolerance and that the similar robustness can be considered.

References


Improvement of Fuzzy ARTMAP by Controlling Match Tracking

Takeshi KAMIO†, Kentaro NOMURA†, Kenji MORI†, Hisato FUJISAKA†, and Kazuhisa HAEIWA†

†Hiroshima City University, 3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima-shi, Hiroshima, 731-3194, Japan
Email: kamio@im.hiroshima-cu.ac.jp

Abstract—ARTMAP is a supervised learning system based on the adaptive resonance theory (ART). It can minimize both the number of recognition categories and the predictive error in noiseless environment. On the other hand, it may generate categories unnecessarily in noisy environment. As a result, the recognition performance decreases. The cause is the match tracking (MT) which is an important function of ARTMAP. To overcome the category proliferation problem, we propose a learning method that can control MT appropriately.

1. Introduction

Adaptive resonance theory neural network (ART-NN) is an unsupervised learning system that can generate and grow the recognition categories by comparing the similarity between inputs and memories with the fineness of classification called the vigilance parameter. In contrast, ARTMAP is a supervised learning system that consists of the learning ART-NN (ARTa), the supervising ART-NN (ARTb) and the map field (MF). ARTa and ARTb classify sample data and recognition codes respectively. MF maps each category in ARTa to the corresponding one in ARTb.

In this paper, we discuss fuzzy ARTMAP [1] that can control MT and shows high robustness against noise, and propose a new method to control MT. As shown in Fig.2, ARTMAP consists of the learning ART-NN (ARTa), the supervising ART-NN (ARTb) and the map field (MF). The learning ART-NN (ARTa) receives a sample a∈[0,1]na that contains noise and the corresponding recognition code b∈[0,1]nb respectively.

Fuzzy ARTMAP is a supervised learning system based on the adaptive resonance theory (ART). It can minimize both the number of recognition categories and the predictive error in noiseless environment. On the other hand, it may generate categories unnecessarily in noisy environment. As a result, the recognition performance decreases. The cause is the match tracking (MT) which is an important function of ARTMAP. To overcome the category proliferation problem, we propose a learning method that can control MT appropriately.

2. Fuzzy ARTMAP

Here, we review the fuzzy ARTMAP [1] and AL-SLMAP method that is the combination of the average leaning (AL) [7] and the slow learning of map field weight vector (SLMAP) [5].

First, we explain the behavior of the fuzzy ART-NN. As shown in Fig.1, it has an attentional subsystem (AS) and an orienting subsystem (OS). AS consists of an input layer (F0), a matching layer (F1) and a category layer (F2). If a normalized original input a∈[0,1]n is given to F0, F0 provides F1 with A∈[0,1]2n. It is the complement code of a (i.e., A=[a1,⋯,an,1−a1,⋯,1−an]). After A goes through F1 and the bottom-up weight U∈ℜ2, each F2 neuron j receives a choice strength Tj:

\[ T_j = |A \land U_j| / (\alpha + |U_j|), \quad (j = 1, \cdots, m), \] (1)

where \( (p \land q) = \min(p, q), \) \( |p| = \sum|p|, \) \( \alpha > 0 \) is the choice parameter and \( m \) is the number of F2 neurons. The neuron J with the maximal \( T_j \) is activated in F2. If several neurons have the maximal \( T_j \), the neuron with the minimal index is selected from them. The F2 activity \( y \in \mathbb{R}^m \) satisfies \( y_j = 1 \) and \( \sum y_j = 1 \). The activated F2 neuron \( J \) provides F1 with \( D \in \mathbb{R}^2 \) that is the same as the top-down weight, and then the F1 activity \( x = A \land D \). On the other hand, OS calculates the matching degree \( S_j \) from \( A \) and \( x \):

\[ S_j = |A \land D_j| / |A|. \] (2)

Moreover, OS compares \( S_j \) with the vigilance parameter \( \rho \in [0,1] \). If \( S_j \geq \rho \), the neuron \( J \) is the category for the present input \( a \). If \( S_j < \rho \), the neuron \( J \) is reset.

Next, we explain the behavior of the fuzzy ARTMAP.

As shown in Fig.2, ARTMAP consists of the learning ART-NN (ARTa), the supervising ART-NN (ARTb) and the map field (MF). In the learning period, ARTa and ARTb receive a sample a∈[0,1]na that contains noise and the corresponding recognition code b∈[0,1]nb respectively.
The vigilance parameter of ARTa (i.e., $\rho_a$) is set to its baseline value $\rho_{a0}$ whenever ARTa receives a new sample. If the category of ARTa is designated by the $F^a_j$ neuron $j$, ARTa provides $F^b$ in MF with $W^a_{jb} \in \mathbb{R}^{ab}$ that is the same as the map field weight. If the category of ARTb is designated by the $F^b_j$ neuron $K$, ARTb provides $F^b$ with $y^b \in \mathbb{R}^{ab}$ that satisfies $y^b_i=1$ and $y^b_j=0$. After receiving $W^a_{jb}$ and $y^b$, MF checks the mapping from ARTa to ARTb by the following equation:

$$[x^a] \rightarrow [y^b] \land W^a_{jb} \geq \rho_{ab} \land |y^b|,$$

where $x^a$ is the $F^a_j$ activity and $\rho_{ab} \in [0,1]$ is the vigilance parameter of MF. If Eq.(3) is true, MF judges that the mapping is correct, and vice versa. In the case of an error mapping, MF executes the match tracking (MT). MT resets the $F^a_j$ neuron $J$ by increasing $\rho_a$ as follows:

$$\rho_a = S^a_j + \varepsilon,$$

where $S^a_j$ is the matching degree of the $F^a_j$ neuron $J$ and $\varepsilon$ is an arbitrary small positive value. Therefore, ARTa selects another category or generates a new one after MT. The above processes are iterated unless Eq.(3) is satisfied. This means that MT can correct an error mapping. When the mapping is correct, AL-SLMAP updates weight vectors as follows:

$$D^a_{j}(\text{new}) = (1+c_1)^{-1} A + (1-(1+c_1)^{-1}) D^a_{j}(\text{old}),$$

$$U^a_{j}(\text{new}) = \beta_a (A \wedge D^a_{j}(\text{new})) + (1-\beta_a) U^a_{j}(\text{old}),$$

$$D^b_k(\text{new}) = \beta_a (B \wedge D^b_k(\text{old})) + (1-\beta_a) D^b_k(\text{old}),$$

$$U^b_k(\text{new}) = \beta_b (B \wedge U^b_k(\text{old})) + (1-\beta_b) U^b_k(\text{old}),$$

$$W^a_{jb}\text{new} = \beta_a (y^a \wedge W^a_{jb}(\text{old})) + (1-\beta_a) W^a_{jb}(\text{old}),$$

where $c_1$ is the count of selections of the $F^a_j$ neuron $J$ in the correct mappings. $\beta_a, \beta_b, \beta_a, \beta_b \in (0,1]$ are the learning rates. All the components of weight vectors are initially set to 1. After the supervised learning is finished, ARTa can be used as the recognition system. This is because $W^a_{jb}$ gives the mapping from ARTa to ARTb.

### 3. Control of Match Tracking (MT)

#### 3.1. Influence of MT on Category Spaces

The original MT can correct an error mapping from ARTa to ARTb by increasing the vigilance parameter $\rho_a$ (i.e., fineness of classification in ARTa). However, the increment of $\rho_a$ may generate categories unnecessarily [6], [7]. To clarify the role of the increment of $\rho_a$, we compare the original MT with the restricted MT that resets an $F^a_j$ neuron without increasing $\rho_a$. We call the former MTup and call the latter MTfix.

Here, let us consider that the fuzzy ARTMAP with AL-SLMAP learns two classification problems shown in Fig.3(a) and Fig.4(a). A sample $x \in \mathbb{R}^2$ is randomly selected from the gray areas and the corresponding recognition code $y \in \mathbb{R}^2$ is defined by the area’s label. Concretely, the code of the areas $A_1$~$A_6$ is $[1,0]$ and that of the area $B$ is $[0,1]$. Fig.3(b) and Fig.4(b) show the category spaces constructed by MTup. Fig.3(c) and Fig.4(c) show the category spaces constructed by MTfix. Each category is illustrated by two rectangles. They are defined by $U^a_{j}$ and $D^a_j$. The substantial area of each category expands outside of these two rectangles. Therefore, the boundary is determined by the position and the size of them.

**Example 1** (see Fig.3): MTfix did not generate an exclusive category for each of areas $A_1$~$A_4$; but MTup did. As a result, the number of categories $m_a$ of MTup was larger than that of MTfix. In the case of MTup, the recognition rate for leaning data $R_1$ was 1.0. But $R_2$ of MTfix was 0.85.

**Example 2** (see Fig.4): MTfix did not generate an exclusive category for each of areas $A_1$~$A_4$; but MTup did. As a result, $m_a$ of MTup was larger than that of MTfix. Both MTfix and MTup achieved $R_1=1.0$.

From these results, we can find the following. MTup can execute fine classification and achieve high $R_2$, finally. However, even if there is an existing category that can learn the present input, MTup may generate a new category.
This phenomenon causes the category proliferation. On the other hand, MTfix can grow each category widely enough to restrict the increment of categories. However, if there occurs unreasonable growth of categories, MTfix cannot achieve high \( R_L \) finally. Therefore, if samples with different recognition codes are selected from a small area (e.g., Fig.3(a)), MTup should be used. If samples with the same recognition code are selected from a large area (e.g., Fig.4(a)), MTfix should be used. Since both situations easily occur in noisy environment, it is necessary to control MT to construct category spaces appropriately. This is the basic idea of the learning method proposed in this paper.

3.2. Proposed Learning Method

As mentioned above, it is necessary to control MT for ARTMAP to construct appropriate category spaces in noisy environment. Therefore, we think up a new function to switch between MTup and MTfix according to the learning progress and add it to AL-SLMP. Furthermore, our proposed learning method can delete the \( F_2^a \) neurons that necessarily cause the error recognition in the test period and has the weight updating equation that differs from the original AL-SLMP.

First, we explain the function to switch between MTup and MTfix. Since our learning method basically uses MTfix, it can restrict the increment of categories. However, the recognition rate for leaning data \( R_t \) may keep low. To solve this problem, our leaning method uses MTup during a short period, if \( R_t \) becomes stable. After ARTMAP processes the \( t \)-th sample, \( R_L(t) \) is calculated as follows,

\[
R_L(t) = \frac{r}{P_R},
\]

where \( r \) is the number of times of the correct mapping from ARTa to ARTb in the period \( [t-P_\ast+1, t] \). The stability condition of \( R_L(t) \) is given by,

\[
| R_L(t-P_\ast) - R_L(t) | \leq \kappa,
\]

where \( \kappa \) is a small positive value. If Eq.(7) is satisfied, MTup is used during the period \( [t+1, t+P_\ast+1] \), and then MTfix is used again. It is clear that switching between MTup and MTfix is executed, even if \( R_L(t) \) is high. However, since ARTMAP with high \( R_L(t) \) hardly executes MT, MTup does not cause the category proliferation. Therefore, our learning method can obtain the advantages of MTup and MTfix.

Next, we explain how to delete the \( F_2^a \) neurons. MT is the important function to correct error mappings from ARTa to ARTb. However, there may be neurons that need MT after the learning is finished. Since MT cannot be used in the test period, such neurons necessarily cause the error recognition. To improve the recognition rate for test data \( R_T \), our learning method deletes the \( F_2^a \) neurons that often use MT. The \( F_2^a \) neuron \( j \) is deleted immediately, if it satisfies the following condition:

\[
d_j / N_L \geq D,
\]

where \( d_j \) is the number of times of MT in \( F_2^a \) neuron \( j \). \( N_L \) is the total number of samples, and \( D \) is the reference
value to delete the $F_2^a$ neurons.

Moreover, our learning method modifies one of weight updating equations of the original AL-SLMAP as follows,

$$U_{j}^{new} = (1 + c_j) \cdot (A \cdot D^{new}) + (1 - (1 + c_j)) \cdot U_{j}^{old}. \tag{9}$$

This modification guarantees that $U_j$ converges according as the learning proceeds. Therefore, it can decrease the bad influence of noise on $U_j$.

4. Simulation Results

Simulations have been carried out to demonstrate the effectiveness of our proposed method (PM). For the alphabet character recognition problem, PM is compared with FCSR [1] and AL-SLMAP [7]. The original patterns of the alphabet characters are shown in Fig.5. Each pattern is illustrated by a (7×7)-pixel image. The pixel values are set to 0 for white pixels and 1 for black ones. In the learning period, ARTa and ARTb receive noisy patterns (i.e., sample data) $a \in \mathbb{R}^{7\times7}$ and the corresponding recognition codes be $\mathbb{R}^{7\times7}$ respectively. A noisy pattern $a$ is constructed by inverting some pixels in an original pattern selected randomly. The number of inverted pixels depends on Hamming distance (HD). In a recognition code $b$, one element is set to 1 and the others are set to 0. For instance, the code $b$ corresponding to the character “A” is $[1,0,\ldots,0]$. The number of sample data $N_s$ is 5000. In the test period, ARTa receives noisy patterns (i.e., test data) $a \in \mathbb{R}^{7\times7}$. The number of test data $N_t$ is 100000. We estimate each learning method by the number of generated categories (i.e., $F_2^a$ neurons) $m_a$ in the learning period and the recognition rate for test data $R_T$. These simulations are iterated 10 times and the results are averaged. The parameters of each learning method are as follows. In the case of FCSR, $\alpha_a=0.1$, $\beta_a=0.2$, $\rho_a=0.5$, $\alpha_b=1$, $\beta_b=1$, $\rho_b=1$, and $\rho_a=1$. In the case of AL-SLMAP, $\alpha_a=0.1$, $\beta_a=0.2$, $\rho_a=0.5$, $\alpha_b=1$, $\beta_b=1$, $\rho_b=1$, $\rho_a=0.02$, and $\rho_b=0.75$. They are the same as Ref.[7]. In the case of PM, $\alpha_a=0.1$, $\beta_a=0.5$, $\alpha_b=1$, $\beta_b=0.02$, $\rho_a=0.02$, $\rho_b=1000$, $P_\rho=500$, $P_u=500$, $x=0.02$, and $D=0.02$, which are the same as AL-SLMAP except for $\beta_a$, $P_\rho$, $P_u$, $\kappa$, and $D$.

Fig.6 and Fig.7 show $m_a$ and $R_T$ respectively. The results show the following. As reported in Ref.[7], AL-SLMAP is much superior to FCSR in terms of the category proliferation and the recognition performance. Furthermore, PM is better than AL-SLMAP when the inputs contain a large amount of noise.

5. Conclusions

In this paper, we have investigated the role of the match tracking (MT) in the fuzzy ARTMAP, and then have proposed the learning method that can control MT. Simulation results have shown that our proposed method is better than AL-SLMAP in terms of the category proliferation and the recognition performance when the inputs contain a large amount of noise.

References


Non-autonomous Chaotic Radio Pulse Generator

Nikolay Atanov, Alexander Dmitriev, Elena Efremova and Lev Kuzmin

Institute of Radio-Engineering and Electronics, Russian Academy of Sciences
Mokhovaya St. 11/7, Moscow, 125009, Russia
Email: chaos@cplire.ru

Abstract— In this report a possibility of generating chaotic radio pulses by means of parameter modulation of chaotic oscillator, which authors earlier demonstrated theoretically, is verified experimentally. Computer simulation and experiment show that the proposed method allows to decrease power consumption of the generator in proportion to the duty cycle.

1. Introduction

Chaotic radio pulses are perspective information carrier for ultra-wideband wireless communication systems [1–3].

Chaotic pulse sequence can be obtained by means of deep amplitude modulation of stationary continuous chaotic signal at the output of chaotic source. However in this case only small part of chaotic source power is transformed into the chaotic pulses.

In Refs. [4, 5] authors proposed to generate chaotic pulses by means of external stimulation of dynamic system by regular signal to obtain chaotic oscillations on only a part of external force period and increase, as a result, the efficiency of pulse generation. This idea was verified by means of computer simulation [4, 5]. As a source of chaotic oscillations, chaotic generator with 2.5 degrees of freedom with bipolar transistor as active element was used. External stimulation of the oscillator was realized by means of alternating the voltage supply.

At the next stage it would seem logical to verify the assumptions and results, obtained earlier by the generator simulation, on a more complex generator model that accounts for characteristics of real radio components.

In this report, workability of this approach is verified by means of computer simulation using Advanced Design System (ADS) software and in a physical experiment.

2. Non-autonomous generator model in ADS

Simple mathematic model of the generator is based on Kirchhoff’s laws and is described by a system of differential equations in the case of lumped-component system or by a system of partial differential equations in the case of distributed-parameter system. In this model voltage-current characteristic of nonlinear elements, such as diodes, bipolar or field-effect transistors is described by as simple as possible piecewise-linear, exponential or polynomial function. This allows us to analyze main properties of system’s dynamics. However, when going from simplified mathematical models to electronic devices, a set of problems occur concerned with mismatch of dynamical modes in mathematical model and in real electronic device, that complicates device realization. In order to solve the problem of transition from mathematical model of a generator to electronic device, specialized software for simulation of the electronic circuits, such as SPIEC, Electronic Workbench or Advanced Design System (ADS) can be used. The most powerful is ADS. The range of problems it works with spreads from development of simple devices like filters, amplifiers, oscillators to simulation of communication system as a whole, and includes all phases from development of principles to solution of system level problems. Besides in ADS one can simulate electronic circuits with element parameters provided by their manufacturers and in this way obtain models as close as possible to experimental prototype. ADS allows simulation of microwave electronic devices. It allows to take into account such characteristics as board topology and substrate properties as well. So, using ADS gives chance to reduce the gap between mathematical model and experimental prototype.

In Fig. 1 the scheme of generator is presented. As a whole this scheme coincides with that described in [4, 5], where a mathematical model of generator with 2.5 degrees of freedom consisted of an active element (bipolar transistor) and a passive quadrupole that defined frequency response of the system. The only difference from scheme of [4, 5] is the presence of block capacity $C_{\text{block}} = 100$ pF and resistance $R = 50$ Ohm at the output of the scheme. Block capacity is driven in scheme to cut off zero-frequency component in the signal power spectrum and the output resistance imitates output load in experimental set.

For the bipolar transistor we use BFP620 transistor by Infineon Technologies. Operating range of this transistor spreads up to 70 GHz.

In ADS we used Gummel-Poon model for bipolar transistor simulation. In this model high-frequency nonlinear effects, temperature effects, etc. are taken into account. Besides, the effect of the package was taken into account.

![Fig. 1. Generator circuit.](image-url)
Parameters of Gummel-Poon model and package model can be found in [6].

The quadrupole in the feedback loop consists of a set of passive elements with parameters: \( L_1 = L_2 = 10 \, \text{nH} \), \( C_0 = C_1 = C_2 = 15 \, \text{pF} \), \( R_0 = 200 \, \text{Ohm} \), \( R_1 = R_2 = 15 \, \text{Ohm} \). Collector voltage is set by fixed voltage source \( V_C = 6.4 \, \text{V} \). Emitter voltage is set by alternating source voltage \( V_E \). From now on voltage amplitudes are depicted regardless of sign.

As was shown in [4, 5], in generator model with current-voltage transistor response described by piecewise-linear function, for certain parameter set chaotic oscillations exist for any value of emitter voltage in excess of threshold value \( V_T = 0.75 \, \text{V} \), i.e. in model [4, 5] oscillation mode doesn’t change with increasing emitter voltage, only oscillation amplitude increases. This phenomenon is specific only for model with piecewise-linear current-voltage transistor response. Since response of ADS transistor model is significantly smooth and nonlinear, the dependence of dynamic modes of the generator on the emitter voltage becomes much more complex.

**Fig. 2. Bifurcation diagram of dynamic modes as a function of parameter \( V_E \).**

Bifurcation diagram (fig. 2) demonstrates dependence of oscillation modes on emitter voltage. As can be seen in the figure, for \( V_E < 1.22 \, \text{V} \) there is no oscillations in the system, for \( V_E \) in the range \( 1.22 \, \text{V} < V_E < 1.7 \, \text{V} \) regular oscillations of period 1 are observed, in the range \( 1.7 \, \text{V} < V_E < 2.05 \, \text{V} \) a number of period doubling bifurcations with transition to chaos take place, and for \( 2.05 \, \text{V} < V_E < 2.5 \, \text{V} \) windows of chaotic and periodic oscillations alternate.

3. **Modulation by external signal**

Let us consider an operation mode with \( V_E = 2.2 \, \text{V} \) and \( V_C = 6.4 \, \text{V} \). In this case, the signal at the output of autonomous generator is chaotic oscillations. Waveform and power spectrum of oscillations are represented in fig. 3.

The generator is stimulated by sinusoidal signal so that emitter voltage is changed slowly with respect to characteristic period of generator oscillations. Non-autonomous system generates pulses with sinusoidal envelope, and inner structure of the pulses changes in accordance to dynamic mode evolution shown in bifurcation diagram in fig. 2.

In fig. 4 chaotic pulse stream obtained by harmonic variation of emitter voltage \( V_E \) with amplitude 2.2 V and frequency \( f = 2.5 \, \text{MHz} \) is shown.

**Fig. 3. (a) Waveform and (b) power spectrum of the output signal of autonomous generator at \( V_E = 2.2 \, \text{V} \).**

**Fig. 4. Chaotic radio pulse stream in the case of harmonic stimulation of the generator.**

To avoid effect of dynamic mode alteration over the duration of chaotic radio pulse we will stimulate generator by square-wave video pulses with the amplitude corresponding to the value of emitter voltage related to chaotic mode of the autonomous generator.

Let the emitter voltage source form a sequence of video pulses with amplitude 2.2 V, 100 ns duration and duty cycle \( \frac{1}{4} \). According to this order of emitter voltage alteration a sequence of chaotic radio pulses of 100-ns duration and duty cycle \( \frac{1}{4} \) is formed at the generator output (fig. 5a). In fig. 5b power spectrum of chaotic radio pulse stream is shown. As can be seen, modulation does not noticeably affect the form of the power spectrum.

Power consumption of the generator in stationary mode \( (V_E = 2.2 \, \text{V}) \) is equal to 64 mW, average power of the generated chaotic signal is 0.32 mW. In modulation mode (modulated video pulse amplitude 2.2 V and duty cycle \( \frac{1}{4} \)) power consumption is 17 mW, average generated signal power is 0.08 mW. So, chaotic radio pulse formation by means of varying emitter voltage allows us to reduce power consumption approximately by a factor of 4, i.e. in proportion to the duty cycle.
4. Pulse identity

As was shown in [5], in mathematical model of the generator it is possible to obtain short identical pulses with complex waveform. This fact is extraordinary, because of instability of phase trajectories of chaotic pulses. Identical pulse forming is possible due to the fact that during time intervals between chaotic pulses the trajectory relaxes to the stable equilibrium point so, that each time the trajectory starts at the leading front of the emitter voltage it starts from practically the same initial conditions. In this case, if pulses are short, exponential divergence of unstable phase trajectories for close but different initial conditions has no enough time to take the trajectories away to the distance comparable with the size of chaotic attractor, which would mean loss of pulse identity.

What will happen in the realistic ADS model? In this case the trajectory cannot reach the equilibrium point, because of thermal noise. Is it possible to form identical radio pulses of complex form in this case?

To ascertain, consider beginning fragments of each chaotic pulse presented in fig. 5a. Superposition of beginning parts of chaotic radio pulses is shown in fig. 6. As can be seen, during approximately 20 ns the pulse waveforms coincide. So the ability of the system to generate sequences of short identical radio pulses with complex waveform is confirmed in fig. 7, where 20-ns radio pulses are shown.

5. Generator prototype

Chaotic generator described in section 1 has been accomplished as an experimental prototype. There are two voltage sources in the scheme: the first for collector voltage $V_C$, the second for emitter voltage $V_E$. Values of these voltages determine dynamic mode of the generator. Parameters of components are the same as in ADS generator model.

As was ascertained experimentally, at a constant value of collector voltage $V_C$ the dependence of dynamic modes of the generator on emitter voltage $V_E$ is as follows: in the range $0 \, V < V_E < 1.3 \, V$ there is no oscillations, at $V_E = 1.3 \, V$ regular oscillations of period 1 are excited, in the range $1.3 \, V < V_E < 2.0 \, V$ a set of period doubling bifurcations and transition to chaos are observed, and at $2 \, V < V_E < 2.5 \, V$ chaotic oscillation mode is realized in the system. Note
that the picture of mode evolution qualitatively coincides with bifurcation diagram obtained in ADS (fig. 2).

In the experiment as an operation mode we have taken the mode with $V_C = 6.4 \text{ V}$, $V_E = 2.04 \text{ V}$. The most part of the power of generated chaotic oscillations is in the frequency range 100…500 MHz (fig. 8a).

6. Pulse generation. Experiment

In numerical simulations emitter voltage was varied harmonically or by means of a sequence of square-wave video pulses [4, 5]. In the experiments to be described only square-wave video pulses were used, i.e. with constant collector voltage the emitter was pumped by square-wave pulses of negative polarization and amplitude 2.04 V. At zero voltage on the emitter there is no signal generation which corresponds to intervals between pulses; at voltage 2.04 V the generator is in chaotic oscillation mode.

In the experiments chaotic radio pulse streams of different duty cycle were generated. In fig. 8b power spectrum of chaotic pulse stream with pulse duration 100 ns and duty cycle $\frac{1}{4}$ is shown. Comparison of figs. 8a and 8b shows that power spectrum of the pulse stream remains continuous and its form is similar to that of stationary chaotic signal of the generator. This testifies that on the time interval of pulse generation its characteristics are close to those of stationary chaotic signal, and transient time at the beginning and at the end of the pulse is significantly shorter than the chaotic pulse duration and has no essential effect on the spectrum form. As can be also be seen, average spectral density in pulse generation mode is lower than in the case of stationary chaotic signal. Measurements show that it decreases in proportion to the duty cycle.

In fig. 8b the generator is in chaotic oscillation mode.

In the experiments chaotic radio pulse streams of different duty cycle were generated. In fig. 8b power spectrum of the pulse stream with pulse duration 100 ns and duty cycle $\frac{1}{4}$ is shown. Comparison of figs. 8a and 8b shows that power spectrum of the pulse stream remains continuous and its form is similar to that of stationary chaotic signal of the generator. This testifies that on the time interval of pulse generation its characteristics are close to those of stationary chaotic signal, and transient time at the beginning and at the end of the pulse is significantly shorter than the chaotic pulse duration and has no essential effect on the spectrum form. As can be also be seen, average spectral density in pulse generation mode is lower than in the case of stationary chaotic signal. Measurements show that it decreases in proportion to the duty cycle.

Fig. 9. Waveform of chaotic radio pulse stream: (a) video pulse duration 100 ns, duty cycle $\frac{1}{4}$, (b) video pulse duration 130 ns, duty cycle $\frac{1}{5}$.

In fig. 9a fragment of the output signal waveform in pulse mode is given. The form of the output pulses testifies to their chaotic nature and insignificance of transient process in comparison with pulse duration. Besides, rather good match of the durations and duty cycles of video pulses on emitter and of output chaotic pulses is observed.

Measurements of power consumption show that it decreases in proportional to duty cycle. So, the energy efficiency of the discussed chaotic radio pulse forming method is higher than that of external modulation, and the advantage increases with decreasing duty cycle.

As was mentioned above, a possibility of generating practically identical pulses of complex form was shown in ADS simulation. However, the presence of this effect in experiment is not obvious due to thermal noise and technical fluctuations.

Typical pulse structure generated by the system is shown in fig. 9b. As can be seen, beginning parts of the pulses (approx. 20 ns) really coincide.

7. Conclusions

Calculations in ADS and experiments testify the possibility of efficient forming of chaotic radio pulses by means of external periodic stimulation of chaotic generator. As a result, power consumption decreases in proportion to the duty cycle. Experimental study showed that the discussed method allows to obtain practically identical ultra-wideband pulses of complex form.

Acknowledgments

This report is supported in part by the Russian Foundation for Basic Research (Grant No. 05-02-17667A) and by Analytical Departmental Target Programme «Development of High School Potential» on 2006–2007 of the Ministry of Education and Science.

References

Detection of Information Symbols and Sequence Lengths
Using Suboptimal Receiver for Chaos Shift Keying

Shintaro ARAI† and Yoshifumi NISHIO‡

†Department of Electrical and Electronics Engineering
Tokushima University, 2-1 Minami-Josanjima, Tokushima 770-8506, Japan
Telephone: +81-88-656-7470, Fax: +81-88-656-7471
Email: arai@ee.tokushima-u.ac.jp, nishio@ee.tokushima-u.ac.jp

Abstract—This paper proposes a new detection method of a receiver for Noncoherent Chaos Shift Keying. In this scheme, the receiver detects not only an information symbol but also a chaotic sequence length which used for Chaos Shift Keying. The detection of the chaotic sequence length is performed by the method adapting the suboptimal receiver proposed in our previous research, and it is the very simple method using the feature of chaos. In order to investigate its detection, we carry out computer simulations and observe the performance.

1. Introduction

Recently, a noncoherent receiver for digital communications systems using chaos is studied actively [1]-[4]. Especially, it is attracted to develop a suboptimal noncoherent receiver having the performance similar to the optimal noncoherent receiver.

In the previous research, we have proposed a suboptimal receiver using very simple algorithm. Our method detects symbols from the calculated values of the shortest distance between received signals and a chaotic map [5]. Furthermore, we extended this concept to the distance in $N_d$-dimensional space using $N_d$ successive received signals ($N_d : 3, 4, \ldots$) [6]. As a result, we obtained the best performance for the dimension $N_d$, which is equal to the length of the chaotic sequence $N$. In addition, we confirmed that the performance of this suboptimal receiver became better as $N$ increased. Moreover, in order to use the average energy per bit effectively and to improve the bit error performance, we proposed a Chaos Shift Keying (CSK) transmitter which the chaotic sequence length changes efficiently and obtained a better BER performance than the existing CSK communication system [7]. However, this simulation was carried out by the case which assumed that the chaotic sequence length was known at the receiving side. Hence, the receiver has to detect the chaotic sequence length, i.e., a point where the chaotic sequence length is changed.

In this study, based on the detection method of the information symbol in the suboptimal receiver proposed in our previous research, we experiment with a detection of the point where the chaotic sequence length is changed. We carry out computer simulations and investigate the performance.

2. System Overview

We consider the discrete-time binary CSK communication system, as shown in Fig. 1.

Figure 1: Block diagram of discrete-time binary CSK communication system.

2.1. Transmitter

In the transmitter, a chaotic sequence is generated by a chaotic map. In this study, we use a skew tent map to generate the chaotic sequence. The skew tent map is one of simple chaotic maps, and it is described by Eq. (1)

$$\begin{align*}
x_{k+1} &= \begin{cases} 
2x_k + 1 - a \\
1 + a 
\end{cases} & (-1 \leq x_k \leq a) \\
\begin{cases} 
-2x_k + 1 + a \\
1 - a 
\end{cases} & (a < x_k \leq 1)
\end{align*}
$$

(1)

where $a$ denotes a position of the top of the skew tent map. The information symbol is modulated by CSK using the skew tent map and its reversal map. In other words, if the information symbol “1” is sent, Eq. (1) is used, and if “0” is sent, the reversed function of Eq. (1) is used.

In order to transmit 1-bit information, $N$ chaotic signals are generated, where $N$ is the chaotic sequence length. Therefore, the transmitted signal is denoted by a vector $S = \langle s_1, s_2, \cdots, s_N \rangle$.

In the case of the existing CSK transmitter, the chaotic sequence length is fixed value. In this study, in order to detect the point where the chaotic sequence length is changed, the transmitter changes the chaotic sequence length at random. Here, we consider a very simple model which
switches 2 different chaotic sequence lengths, \( N = 8 \) or \( N = 12 \). Also, the initial value is changed for every information symbol.

2.2. Channel and Noise

In the channel, a noise is assumed to be additive white Gaussian noise (AWGN) and is denoted by the noise vector \( n = (n_1 \ n_2 \ \cdots \ n_N) \). Thus, the received signal vector is given by \( R = (R_1 \ R_2 \ \cdots \ R_N) = S + n \).

2.3. Receiver

The receiver recovers the transmitted signals from the received signals and demodulates the information symbol. In this study, we use the suboptimal receiver proposed in our previous research [6]. Based on the detection method of the information symbol in the suboptimal receiver, we experiment with the detection of the point where the chaotic sequence length is switched.

2.4. Detection Method of Information Symbol [6]

![Detection method](image)

The suboptimal receiver proposed by the authors calculates the shortest distance between received signals and the map in the \( N_d \)-dimensional space using \( N_d \) successive received signals (\( N_d : 3, 4, \cdots \)).

As an example, we explain the case of \( N_d = 3 \). Figure 2 shows the 3-dimensional space of the skew tent map whose coordinates correspond to the three successive received signals \( (R_k, R_{k+1}, R_{k+2}) \) where \( k = 1, 2, \cdots, N - 2 \). In order to decide which map is closer to the point \( (R_k, R_{k+1}, R_{k+2}) \) in the 3-dimensional space in Fig. 2, the shortest distance between the point and the map has to be calculated. Therefore, we calculate the shortest distance using the scalar product of the vector.

Any two points of \((x_0, y_0, z_0)\) and \((x_1, y_1, z_1)\) are chosen from each straight line in the space of Fig. 2, as shown in Fig. 3. In Fig. 3, a unit vector \( u \) is calculated from \((x_0, y_0, z_0)\) and \((x_1, y_1, z_1)\) by the following equation,

\[
u = (l, m, n) = \left( \frac{x_1 - x_0}{A}, \frac{y_1 - y_0}{A}, \frac{z_1 - z_0}{A} \right) \tag{2}\]

where \( A = \sqrt{(x_1 - x_0)^2 + (y_1 - y_0)^2 + (z_1 - z_0)^2} \). In addition, vector \( v_0 \) is calculated from \((R_k, R_{k+1}, R_{k+2})\) and \((x_0, y_0, z_0)\) by the following equation.

\[
v_0 = (R_k - x_0, R_{k+1} - y_0, R_{k+2} - z_0) \tag{3}\]

Product \( T \) in \( u \) and \( v_0 \) is calculated by the following equation.

\[
T = l(R_k - x_0) + m(R_{k+1} - y_0) + n(R_{k+2} - z_0) \tag{4}
\]

Hence, \( v_1 \) can be calculated from the product of \( T \) and \( u \). Therefore, we can calculate the point with the shortest distance \( (X, Y, Z) \) and the shortest distance \( D \) by the following equations.

\[
(X, Y, Z) = (Tl + x_0, Tm + y_0, Tn + z_0) \tag{5}
\]

\[
D = \sqrt{(X - R_k)^2 + (Y - R_{k+1})^2 + (Z - R_{k+2})^2} \tag{6}
\]

Note that if the point is outside the cube, we calculate the distance between the point and the nearest edges of the maps.

For the 3-dimensional case, there are four straight lines in the space. Therefore, the minimum value in four distances is chosen as the shortest distance \( D_1 \) for symbol “1”. In the same way, \( D \) of symbol “0” is chosen as \( D_0 \). We calculate both of \( D_1 \) and \( D_0 \) for all \( k \) and find their summations \( \sum D_1 \) and \( \sum D_0 \). Finally, we decide the decoded symbol as 1 (or 0) for \( \sum D_1 < \sum D_0 \) (or \( \sum D_1 > \sum D_0 \)).

The calculation of the shortest distance can be extended to \( N_d \)-dimensional space for \( N_d \geq 4 \).

3. Detection Method of Chaotic Sequence Length (Proposed Method)

Based on the detection method of the information symbol in the suboptimal receiver, we explain the detection method of the point where the chaotic sequence length is switched.

Figure 4(a) shows a time series where the chaotic sequence is switched at \( k = 12 \). This sequence contains two chaotic sequences of the information symbols “1” and “0”. We assume that the noise is not added to this sequence. Figure 4(b) shows the calculation results of the shortest distances for every \( k \). Here, we use 4-dimensional space for...
the calculation of the shortest distances. In Fig. 4(b), the solid line and the dotted line show $D_1$ and $D_0$, respectively. Since the noise is not added, the difference between the two lines is very large for $k = 1 \sim 9$. However, the difference between the two lines becomes small for $k = 10 \sim 12$. In order to observe the difference in detail, we show the absolute value of the difference between $D_1$ and $D_0$, as Fig. 4(c). We can see that the differences around the switching point ($k = 12$) are much smaller than other differences. In order to expound on this cause, we use Fig. 5.

Figure 5: Cause of small difference between two lines around switching point.

Figure 5 shows the switching of two chaotic sequences. We can observe that the two chaotic sequences are mixed for the calculation around the switching point ($k = 12$). Therefore, the shortest distance around the switching point is calculated as if the noise was added. This is the key point of the proposed method. Namely, if the difference between $D_1$ and $D_0$ around the switching point becomes very small, it becomes easy to find the point where the chaotic sequence length is switched. In this study, we consider a very simple model which switches 2 values ($N = 8$ and $N = 12$). Hence, the receiver calculates 2 differences between $D_1$ and $D_0$ around 8 and 12, and these two values are compared. Finally, the receiver can decide the chaotic sequence length by which is smaller between $N = 8$ and $N = 12$. In the case of Fig. 4(c), the difference around $N = 12$ is smaller than that around $N = 8$. Therefore, the chaotic sequence length is decided as $N = 12$. Moreover, since $D_1$ is smaller than $D_0$, the information symbol is detected as “1”.

In the same way, the proposed method can detect the chaotic sequence length and the information symbol even when the same information symbol continues like “1” and “1”, as Fig. 6. From Figs. 6(b) and (c), we can find that the chaotic sequence length and the information symbol is detected as $N = 8$ and “1”, respectively.

This detection method improves the performance when $D_1$ and $D_0$ around the switching point is very small. Namely, the initial value should be chosen such that the distances in the $N_q$-dimensional space between the signal point and the two maps, which correspond to the symbols “1” and “0”, is very small. Therefore, when choosing the initial value of the next transmitted signal, the transmitter calculates $D_1$ and $D_0$ using two values, the last value of the
previous transmitted signal and the random value. This operation is performed repeatedly. Finally, the random value which is the closest to \( D_1 \) and \( D_0 \) is chosen as the initial value of the next transmitted signal.

4. Simulation Result

In this section, we study the performance of the proposed method by computer simulations. The simulation conditions are as follows.

In the transmitting side, the chaotic sequence length \( N \) is switched as \( N = 8 \) or \( N = 12 \). In the channel, noise is assumed to be only AWGN. Hence, the noise at the transmitter and the receiver are not considered. In order to calculate the shortest distance, we use the 4-dimensional and 8-dimensional space on the receiving side. Based on these conditions, the system performance is evaluated by plotting the sequence length error rate and BER against \( E_b/N_0 \) when \( 10^4 \) bits of information are transmitted. Here, we assumed that the bit error rate (BER) is calculated in the case where the chaotic sequence length is detected correctly.

In this study, based on the detection method of the information symbol in the suboptimal receiver proposed in our previous research, we experimented with a detection of the chaotic sequence length, i.e., the point where the chaotic sequence length is changed. We have confirmed that the detection of the information from the chaotic sequence is possible even when the chaotic sequence length changed. Moreover, in the previous research, we proposed the CSK transmitter which the chaotic sequence length changes efficiently and obtained a better BER performance than the existing CSK communication system [7]. Therefore, investigating the performance of the chaos communication system which changes the chaotic sequence length efficiently using the proposed method in this study is our future work.

5. Conclusions

In the transmitting side, the chaotic sequence length \( N \) is switched as \( N = 8 \) or \( N = 12 \). In the channel, noise is assumed to be only AWGN. Hence, the noise at the transmitter and the receiver are not considered. In order to calculate the shortest distance, we use the 4-dimensional and 8-dimensional space on the receiving side. Based on these conditions, the system performance is evaluated by plotting the sequence length error rate and BER against \( E_b/N_0 \) when \( 10^4 \) bits of information are transmitted. Here, we assumed that the bit error rate (BER) is calculated in the case where the chaotic sequence length is detected correctly.

In this section, we study the performance of the proposed method by computer simulations. The simulation conditions are as follows.

In the transmitting side, the chaotic sequence length \( N \) is switched as \( N = 8 \) or \( N = 12 \). In the channel, noise is assumed to be only AWGN. Hence, the noise at the transmitter and the receiver are not considered. In order to calculate the shortest distance, we use the 4-dimensional and 8-dimensional space on the receiving side. Based on these conditions, the system performance is evaluated by plotting the sequence length error rate and BER against \( E_b/N_0 \) when \( 10^4 \) bits of information are transmitted. Here, we assumed that the bit error rate (BER) is calculated in the case where the chaotic sequence length is detected correctly.

In this study, based on the detection method of the information symbol in the suboptimal receiver proposed in our previous research, we experimented with a detection of the chaotic sequence length, i.e., the point where the chaotic sequence length is changed. We have confirmed that the detection of the information from the chaotic sequence is possible even when the chaotic sequence length changed. Moreover, in the previous research, we proposed the CSK transmitter which the chaotic sequence length changes efficiently and obtained a better BER performance than the existing CSK communication system [7]. Therefore, investigating the performance of the chaos communication system which changes the chaotic sequence length efficiently using the proposed method in this study is our future work.

References


Yutaka Jitsumatsu†, Masato Ogata‡ and Tohru Kohda†

†Dept. of Computer Science and Communication Engineering, Kyushu University
6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan
‡Dept. of Electrical Engineering, Kyushu Sangyo University
2-3-1 Shokadai, Higashi-ku, Fukuoka, 813-8503, Japan

{jitumatu, kohda}@csce.kyushu-u.ac.jp, mogata@te.kyusan-u.ac.jp

Abstract—Band-limited direct sequence/code division multiple access (DS/CDMA) system employs pulse shaping filters, which are generally implemented by FIR filters. Inter-symbol interference (ISI)-free condition on pulse shaping filters is relaxed. Using Gaussian FIR filter in conjunction with Markovian spreading codes shows better performance than root raised cosine filter with i.i.d. codes.

1. Introduction

Spread spectrum (SS) codes generated by a Markov chain can mitigate multiple-access interference (MAI) and give better bit error rate (BER) performance than independent and identically distributed (i.i.d.) random codes and linear feedback shift register (LFSR) codes in chip-asynchronous direct sequence/code division multiple access (DS/CDMA) systems [1–3]. However, in [1–3] chip waveforms were assumed to be rectangular, which occupy infinite bandwidth. In many commercial systems, root raised cosine (RRC) pulse shaping with excess bandwidth 0.22 is employed. In case of RRC pulse, the improvement given by Markov codes is considerably reduced [4,5]. However, it was also shown that RRC pulse is suboptimal in the presence of white Gaussian noise, the optimum receiver power is perfectly controlled. The received signal is given by

\[ r(t) = \sum_{j=1}^{K} d(j)(t) \ast x(j)(t) \ast hT(t) \ast \delta(t - t_j) + n_0(t), \]

where \( hT(t) \) is the pulse response (FIR) digital filters. Such FIR filters are generally implemented as finite impulse response (FIR) digital filters. Such FIR filters are not band-limited in a strict sense. Hence small amount of out-of-band energy has to be allowed. In addition to the bandwidth requirement, chip waveform usually guarantees zero inter-symbol interference (ISI). The ISI-free condition is of crucial importance in frequency division multiple access (FDMA) system and time division multiple access (TDMA) systems. However, in CDMA systems signal degradation is not mainly caused by ISI but multiple-access interference (MAI) and background noise. Thus, we can relax the ISI-free condition. In [6, 7] BER performance was shown to be improved by ISI pulse.

We already examined Gaussian pulse as a typical example of nonzero ISI pulse and reported that Gaussian pulse with the same excess band energy as RRC pulse gives lower BER than RRC pulse in conjunction with Markovian codes [8]. This paper considers a FIR pulse shaping filter design for asynchronous CDMA systems. Quality of FIR filter is affected by sampling frequency, the number of filter taps and quantization level of filter coefficients. These three parameters of Gaussian and RRC filters as well as BER performance are analyzed, which shows the superiority of Markovian SS codes.

2. Chip-asynchronous DS/CDMA systems

Let us consider a baseband equivalent model of chip-asynchronous DS/CDMA system with \( K \) users. Data and chip durations are \( T_d \) and \( T_c = T_d/N \), where \( N \) is the spreading factor. Data and SS code signals for \( i \)-th user are, respectively,

\[ d(i,t) = \sum_{p=-\infty}^{\infty} d_p(i) \delta(t - pT_d), \]

\[ x(i,t) = \sum_{n=0}^{N-1} x_n(i) \delta(t - nT_c), \]

where \( d_p(i) \) and \( x_n(i) \) are the data symbol and spreading code assumed to be antipodal binary.

Suppose that transmitter and receiver employ pulse shaping filters, denoted by \( h_T(t) \) and \( h_d(t) \), respectively. They are common for all users and normalized to be \( \int_{-\infty}^{\infty} |h_T(t)|^2 dt = \int_{-\infty}^{\infty} |h_d(t)|^2 dt = 1 \). The impulse response of the cascade of the two filters is denoted by \( h_3(t) \). In the presence of white Gaussian noise, the optimum receiver filter is \( h_d(t) = h_T(\bar{z} t) \), where \( \bar{z} \) denotes the complex conjugate of \( z \). In the presence of interference which is not Gaussian process, using a pulse shape other than \( h_T(\bar{z} t) \) minimizes bit error rate [9, 10]. The \( k \)-th user’s propagation delay is \( 0 \leq t_k < T_d \). Assume that the received signal power is perfectly controlled. The received signal is given by

\[ r(t) = \sum_{j=1}^{K} d(j)(t) \ast x(j)(t) \ast h_T(\bar{z} t) \ast \delta(t - t_j) + n_0(t), \]
where an asterisk sign denotes a convolution and \( n_0(t) \) is an additive white Gaussian noise with two-sided spectral density \( \frac{1}{2\pi} \).

Define the Pursley’s aperiodic cross-correlation function [11] as

\[
[X^{(i)} \ast X^{(j)}]_\ell = \begin{cases} \sum_{n=0}^{N-\ell-1} X^{(i)}_n X^{(j)}_{n+\ell} & \text{for } \ell \geq 0, \\ \sum_{n=0}^{N-1} X^{(i)}_n X^{(j)}_{n+\ell} & \text{for } \ell < 0, \end{cases}
\]

(4)

where \( \ast \) denotes the convolution of two sequences. Let \( \text{MAI}^{(i,j)}(t) = x^{(i)}(t) \ast x^{(j)}(-t) \ast h_N(t) = \sum_{n=1}^{N-1} [x^{(i)} \ast x^{(j)}]_k \cdot h_N(t - kT_c) \). Then the correlator output of the 0-th bit is

\[
z^{(i)}(0) = d_0^{(i)} \cdot \text{MAI}^{(i,0)}(0) + \sum_{p=1}^{K} d_p^{(i)} \cdot \text{MAI}^{(i,-pT_d)} + \eta^{(i)}(0),
\]

(5)

where the first, second, third and fourth terms of the right hand side are the signal component, ISI, MAI, and noise component, respectively. The transmitted data symbol is estimated as \( \hat{d}_0^{(i)} = \text{sgn}(z^{(i)}(0)) \).

The bit error rate of \( i \)-th user can be estimated as

\[
\bar{P}_e^{(i)} = Q \left( \frac{\sqrt{\frac{1}{\sigma_N^2 + [\text{MAI}^{(i,0)}(0)/N]}^2}}{2\sigma_N^2 + [\text{MAI}^{(i,0)}(0)/N]} \right),
\]

(6)

where \( Q(x) = \frac{1}{\sqrt{\pi}} \int_x^\infty \exp(-u^2/2) du \) and

\[
\sigma_N^2 = \frac{N_0}{2N} \sum_{k=1}^{N-1} [x^{(i)} \ast x^{(j)}]_k \cdot h_R \ast h_R(kT_c),
\]

(7)

\[
[\sigma_{\text{ISI}}^{(i,j)}]^2 = \frac{1}{N} \sum_{p=0}^{N-1} \text{MAI}^{(i,j)}(-pT_d)^2
\]

(8)

\[
[\sigma_{\text{MAI}}^{(i,j)}]^2 = \text{Var} \left[ \frac{1}{\sqrt{N}} \sum_{p=0}^{N-1} d_p^{(i)} \text{MAI}^{(i,j)}(t_j - t_j - pT_d) \right] \]

\[
= \frac{1}{N} \sum_{p=0}^{N-1} \text{MAI}^{(i,j)}(t_j - t_j - pT_d)^2.
\]

(9)

Since \( j \)-th user’s time delay is not known to \( i \)-th user, we take the expectation on \( [\sigma_{\text{MAI}}^{(i,j)}]^2 \) with respect to the time delay. We obtain

\[
\frac{1}{T_d} \int_0^{T_d} [\sigma_{\text{MAI}}^{(i,j)}]^2 dt_j
\]

\[
= \frac{1}{NT_d} \sum_{k=2}^{2N-2} [x^{(i)} \ast x^{(j)} \ast x^{(j)} \ast x^{(j)}]_k \cdot h_N \ast h_N(kT_c)
\]

(10)

3. Markovian SS Codes

Let \( X \) and \( Y \) be random variables generated by an irreducible aperiodic stationary Markov chain with transition probability matrix

\[
P(A) = \begin{pmatrix} 1 + A & 1 - A \\ 1 - A & 1 + A \end{pmatrix}.
\]

(11)

This transition matrix gives a stationary distribution \((\frac{1}{2}, \frac{1}{2})\) and eigenvalues \((1, \lambda)\). For \( \ell, m, k \geq 0 \), we have

\[
E_X[N_0] = 0, \ E_{XY}[X_n Y_m] = 0
\]

(12)

\[
E_X[X_{n+1} Y_{m+1}] = \lambda^\ell,
\]

(13)

\[
E_X[X_{n+1} X_{n+1} Y_{m+1} X_{n+1} X_{n+1} Y_{m+1}] = \lambda^{\ell+m},
\]

(14)

\[
E_X[N_0] = 0, \ E_{XY}[X_n Y_m] = 0
\]

(15)

where \( E_Z[\cdot] \) denotes the expected value with respect to the distribution of a random variable \( Z \) [3, 12, 13]. Eqs. (12)–(15) give

\[
\frac{1}{N} E_X[|X \ast X|] = \lambda^{\ell},
\]

\[
\frac{1}{N} E_X[|X \ast X \ast Y \ast Y|] = \left( |\ell| + \frac{1 + \lambda^2}{1 - \lambda^2} \right) \lambda^{\ell},
\]

(16)

as \( N \to \infty \). See [14] for derivation and expression for finite \( N \).

4. Pulse Shaping Filter

In this paper, incompletely band-limited pulse shaping filters are considered. Moreover, we relax the ISI-free condition on pulse shaping filter to minimize bit error rate. The variances of interferences and noise are analyzed. This paper considers only a case that the pulse shaping filter \( h_T(t) \) and the chip-matched filter \( h_R(t) \) are the same.

4.1. Root Raised Cosine Pulse

For a root raised cosine pulse, we have

\[
h_T(t; \beta) = \frac{1}{\sqrt{T_c}} \times \frac{4\beta \cos((1 + \beta)\pi t/T_c) + T_c/(1 - \beta)\pi \sin((1 - \beta)\pi t/T_c)}{1 - (\frac{\pi t}{T_c})^2},
\]

(17)

\[
h_R(t) = h_T(t) \ast h_R(t) = \text{sinc}(t/T_c) \cos(\pi(\beta t/T_c))/1 - 4(\beta t/T_c)^2,
\]

(18)

where \( \beta \) is the roll-off factor. Then, for integer delay

\[
h_R \ast h_R(kT_c) = \begin{cases} 1 - \beta/4 & \text{if } k = 0 \\ \frac{\sin(\pi(k+1)/(2\pi T_c))}{\sin(\pi k/(2\pi T_c))} \cdot \frac{(-1)^{k+1}}{\sin(\pi k/(2\pi T_c))} & \text{if } k = \pm 1/\beta \\ \text{otherwise} \end{cases}
\]

(19)

Many practical systems employ nonzero excess bandwidth: typically it is \( \beta = 0.22 \). For RRC pulses, MAI reduction ratio by Markovian codes is only approximately 3% [8]. In order to enhance the variance reduction effect, function of \( h_N \ast h_N(k \cdot T_c) \) for \( k \neq 0 \) should take large value.

\footnote{Note that this asymptotic behavior is not in terms of the number of users \( K \) but in terms of code length \( N \). On the other hand, standard Gaussian approximation (SGA) [11] is valid only for a large \( K \).}
4.2. Gaussian Pulse

Define a Gaussian pulse with a parameter $\alpha > 0$ as

$$h_T(t; \alpha) = \sqrt{\frac{\alpha}{T_c}} \text{gauss}\left(\frac{at}{T_c}\right),$$

(20)

where $\text{gauss}(x) = \sqrt{\frac{2}{\pi}} \exp(-\pi x^2)$. The energy of this pulse is concentrated with 92.4% both in time domain $t \in [-T_c/2 : T_c/2]$ and in frequency domain $\omega \in [-\pi/T_c : \pi/T_c]$ when $\alpha = 1$. We obtain

$$h_N(t) = \frac{1}{\sqrt{2}} \text{gauss}(at/\sqrt{2}T_c), \quad (21)$$

$$h_N(t) * h_N(t) = T_c \frac{1}{\sqrt{2} \alpha} \text{gauss}(at/2T_c). \quad (22)$$

We set $\alpha = 0.863$ for a Gaussian chip pulse in order to make the excess band energies as same as a RRC pulse with $\beta = 0.22$. The excess band energies for both pulses are 4.0%, i.e.

$$\int_{-\pi/T_c}^{\pi/T_c} |H_T(\omega)|^2 d\omega \int_{-\infty}^{\infty} |H_T(\omega)|^2 d\omega = 0.960, \quad (23)$$

where $H_T(\omega)$ is the Fourier transform of $h_T(t)$. Fig. 3 shows BERs for root raised cosine and Gaussian filters. Gaussian chip pulse sacrifices ISI, however, they give lower BER than RRC pulses.

4.3. FIR Filter Design

For digital communications through band-limited channels, pulse shaping filters are employed. It is widely accepted to assume pulse shaping filters are purely bandlimited. However, such filters have infinite time response from minus infinity to plus infinity; therefore they do not satisfy causality. Practically, pulse shaping filters are implemented as finite impulse response (FIR) digital filters. Thus, we consider that pulse shaping filters are not completely band-limited but truncated with a proper time period.

For FIR pulse shaping filter, four filter design specifications are important: the passband edge, the stopband edge, the passband ripple, and the stopband ripple. Parks-McClellan algorithm is widely used FIR filter design method. An improved FIR filter design with equi-ripple is presented in [15]. Quality of the filter depends upon the sampling frequency, the number of filter taps, and quantization level of the filter coefficients. Another major FIR filter design is the windowing method, i.e. impulse response of the ideal low-pass filter is multiplied by the window function, such as Hanning, Hamming, Blackman, Kaiser windows. The choice of window types as well as their design parameters depends on the required quality of FIR filters.

In this paper, FIR filter coefficients are obtained by MATLAB programming [16] with the number of filter taps

![Figure 1: Impulse response of the cascade of the two FIR filters: the number of taps for each filter is 31 and oversampling factor is 4.](image)

![Figure 2: Frequency response of Gaussian and root raised cosine filters](image)
Figure 3: Bit error rates for root raised cosine ($\beta = 0.22$), Gaussian ($\alpha = 0.863$), and rectangular pulses with Markovian codes and i.i.d. codes against signal to noise ratio (SNR) level $2 - 18$ [dB], where the number of users is $K = 10$ and the spreading factor $N = 63$.

$L = 31$ and oversampling rate $F_s = 4$. Fig. 1 shows impulse responses of the cascade of the transmitter filter with the matched filter at the receiver. Frequency responses are illustrated in Fig. 2. Frequency response of the RRC FIR filter has ripples of around -40dB in the stopband. On the other hand, frequency response of Gaussian FIR filter is far from flat spectrum in the passband region (normalized frequency $0 \leq f < 0.125$) but it is less than -40dB at $f = 0.25$ and decreases rapidly for higher frequency.

5. Conclusion

We analyzed the performance of Gaussian pulse which are realized by FIR filters in an asynchronous DS-CDMA system. Although Gaussian pulses have non-zero ISI, they show better BER performance than the conventional ISI-free pulse if it is used in conjunction with Markov codes. Note that Markovian SS codes should be jointly optimized together with pulse shaping filters.

References


Shannon Capacity of Chaos-Based and Conventional Asynchronous Multi-Code DS-CDMA Systems over AWGN Channels

Stefano Vitali†, Riccardo Rovatti† and Gianluca Setti‡

†University of Bologna – ARCES, via Toffano 2/2, 40125 Bologna, Italy
‡University of Ferrara – ENDIF, via Saragat 1, 44100 Ferrara, Italy
E-mail: {svitali, rovatti}@arces.unibo.it, gsetti@ing.unife.it

Abstract— Theoretical limits of asynchronous multi-code DS-CDMA systems over AWGN channels are investigated by computer simulations using the statistics of the associated Shannon capacity. The system capacity is identified with the maximum rate at which an errorless link may operate and is a random quantity depending on the spreading sequences and on the users relative delays and phases. A comparison is reported between chaos-based spreading, possibly complemented by an off-line orthogonalization method and OVSF spreading with a subsequent on-line scrambling in order to address the trade-off between the orthogonality of different (but synchronous) codes of the same user and the auto-correlation properties that counter asynchronous multiple-access interference. We here report that chaos-based strategies result in a non negligible increase in capacity with respect to the classical approach.

1. Introduction

Within the past decade, it has become possible to directly approach the study of chaotic systems and to unravel their potential in several common signal processing tasks. The use of periodically repeated slices of chaotic trajectories as spreading sequences in DS-CDMA systems has attracted attention due to the possibility of showing that it may be used to achieve the minimum possible multiple access interference perceived by a single-user receiver (see e.g. [1][2]). Furthermore, [3] shows that chaos-based spreading improves performance of multi-code DS-CDMA with single-stream correlation-based receiver when the number of synchronous streams per user is relatively low. Though improvements have also been demonstrated in more complex situations (see e.g. [4]), a more general point of view has been recently adopted by some authors. Instead of addressing a particular receiver structure, the intrinsic capability of the transmission mechanism is evaluated from the point of view of its ultimate limits as expressed by the Shannon capacity concept.

We here aim at continuing the work in [5] by extending it to the multi-code DS-CDMA system, interpreting the spreading operation as part of an AWGN vector channel and exploiting the concept of channel capacity as defined by Shannon to compare chaos-based strategies with the classical method employing Orthogonal Variable Spreading Factor (OVSF) sequences and a subsequent scrambling (we will indicate this approach as OVSF+S). This reference approach is, for example, employed in the UMTS standard.

The paper is organized as follows: section 2 introduces the multi-code DS-CDMA system and the capacity formula we have exploited in the Monte-Carlo simulations, section 3 briefly recalls the OVSF+S method and discuss the chaos-based spreading and related strategies while sections 4 and 5 report simulation setup and results respectively.

2. System model

In a multi-code DS-CDMA system any of the U users that access the communication channel at the same time is allowed to transmit in parallel M streams of information bits using a different spreading code per stream. Spreading sequences are generated once and stored in the transmitter and in the receiver. Figure 1 shows a baseband multi-code DS-CDMA multi-user transmitter where indexes u and m identify the generic user and the generic stream respectively. Each
stream is also associated with a spreading sequence $y_{m,n}^u \in \{-1/\sqrt{N}, 1/\sqrt{N}\}$.

Stretches of $N$ symbols of this spreading sequence are used by the bits before transmission. In particular, the $j$-th bit $b_j^{m,n}$ is transmitted as $b_j^{m,n} s_j^{m,n}(t)$ with $s_j^{m,n}(t) = T^{-1/2} \sum_{k=0}^{N-1} y_{j+k}^{m,n} e(t-(jN+k)T)$, where $g(t)$ is the rectangular pulse of unit amplitude and duration $T$. Hence, the $n$-th transmitter produces the signal $\sum_{m=0}^\infty \sum_{j=0}^{N-1} b_j^{m,n} s_j^{m,n}(t)$ that is sent over the channel and arrives at the receiver with a delay $t^r \in [0, NT]$ taking its parts in the total received signal $r(t) = \sum_{m=0}^\infty \sum_{j=0}^{N-1} b_j^{m,n} s_j^{m,n}(t-t^r) + v(t)$ where $v(t)$ is the unavoidable WGN with power $\sigma^2$.

Delay dynamics will be assumed to be much slower than bit rate to fit within the classical information theory view and delays will be considered constant for all the transmission and uniformly distributed in $[0, NT]$.

Decoding operation is performed in the time window $\mathcal{T} = [0, BN_T]$, i.e. for $B$ bit durations. Due to the asynchronous transmission, in general $B + 1$ bits of each user overlap with this window.

If $\phi_j \ (i = 0, \ldots, \infty)$ is an orthonormal basis in $W$ we may think that the decoding process depends on the projections $r_i = \langle r(t), \phi_i(t) \rangle = \int_0^{BN_T} r(t) \phi_i(t) dt$ of $r(t)$ on such a basis. Consider $n_\phi$ projections arranged in the vector $\mathbf{r} = [r_0, \ldots, r_{n_\phi-1}]^T$. They are linked to the vector of transmitted bits $\mathbf{b} = [b_j^0, \ldots, b_j^0, b_j^1, \ldots, b_j^1, b_j^2, \ldots, b_j^2, b_j^{M-1}, \ldots, b_j^{M-1}]$ by $\mathbf{r} = \mathbf{H} b + \nu$ where $\nu$ is the vector of noise projections and $\mathbf{H}$ can be expressed by defining the matrices $\mathbf{H}^{(k,m)}$ whose elements are $H_{q_{m,n}}^{(k,m)} = s_j^{m,n}(t-t^r)$ and writing

$$\mathbf{r} = \left[ H^{(0,0)} \cdots H^{(0,M-1)} \cdots H^{(M-1,M-1)} \right] \mathbf{b} + \nu$$

The above discussion highlights that the system is linear from the transmitted bits to the received projections.

We know from [6] that the Shannon capacity of a vector transmission system characterized by unit energy bits, by a transfer matrix $\mathbf{H}$, and subject to a noise of power $\sigma^2$ is given by

$$C = \log_2 \det \left( I_{n_\phi} + \frac{\mathbf{HH}^T}{\sigma^2} \right)$$

where $I_{n_\phi}$ is the $n_\phi \times n_\phi$ identity matrix, $\mathbf{H}^T$ is the transpose conjugate of $\mathbf{H}$ and $K = U M (B + 1)$.

Note that the two expressions are equivalent because the non-vanishing eigenvalues of $\mathbf{HH}^T$ and $\mathbf{H} \mathbf{H}^T$ are the same as it can be easily checked considering that for any eigenvector $\mathbf{x}$ of $\mathbf{HH}^T$, $\mathbf{y} = \mathbf{H} \mathbf{x}$ is an eigenvector of $\mathbf{H} \mathbf{H}^T$ with the same eigenvalue. This identity allows us to highlight the importance of the correlation between the signals used as spreading waveforms. Equation (1) is commonly adopted to evaluate Shannon Capacity of DS-CDMA systems (see e.g. [7]).

Formally, we may exploit the above defined $\mathbf{H}^{(m,n)}$ to express the matrix $\mathbf{H}^T \mathbf{H}$ as in Table 1. The building block of such a matrix has entries

$$((\mathbf{H}^{(m,n)} \mathbf{H}^{(m',n')}))_{p,q} = \sum_{t=0}^{n_\phi} \left( \sum_{t=0}^{n_\phi} \sum_{t=0}^{n_\phi} \right)$$

We may then let $n_\phi \to \infty$ and assume that the functions $\phi_i$ form a complete basis to obtain

$$((\mathbf{H}^{(m,n)} \mathbf{H}^{(m',n')})_{p,q} = \int_{0}^{t} (\mathbf{s}_q^{m,n}(t-t^r)) \mathbf{s}_q^{m',n'}(t-t^r) dt$$

revealing how $\mathbf{H} \mathbf{H}^T$ is made of all the possible correlations between the spreading waveforms associated to the bits.

Note that the diagonal elements of $\mathbf{H} \mathbf{H}^T$ (that are the diagonal elements of the submatrices $(\mathbf{H}^{(m,n)})^2$) are nothing but the fraction of energy of the $(a,m)$-th spreading waveform that falls within the decoding window and are independent of the spreading sequences. Since the matrix $\mathbf{K} + \mathbf{H} \mathbf{H}^T/\sigma^2$ is Hermitian, its determinant is bounded from above by the product of its diagonal entries [8]. Hence, capacity is maximized when $\mathbf{H} \mathbf{H}^T$ is as much diagonal as possible and the role of sequence design is that of making off-diagonal entries as small as possible.

This hints at a fundamental trade-off between orthogonality of the synchronous waveforms (those corresponding to streams of the same user) and the ability of sequences of countering cross-correlation with random time-shifts.

### 3.3. Spreading strategies

The classical OVSF+1 approach employs orthogonal waveforms and scrambles them to achieve some immunity to asynchronous interference. The chaos-based approach we propose and analyze here starts from sequence generators that are known to minimize (in the average) asynchronous cross-correlation and tries to add orthogonality with a possible subsequent step.

The OVSF+S technique starts from the consideration that if $M \leq N$ and $N$ is power of 2 ($N$ integer multiple of 4 would be enough if one resorts to Hadarmad construction) then there is a set of sequences $\mathbf{w}_u^{m,n}$ in $\{-1, +1\}$, for $m = 0, \ldots, M - 1$ and $u = 0, \ldots, N - 1$ that are pairwise orthogonal [9]. It then assumes that $U$ scrambling codes $\mathbf{w}_u^{m,n}$ $u = 0, \ldots, U - 1$ and $k = -\infty, \ldots, \infty$ can be generated as sequences of independent and uniformly distributed antipodal symbols (this is usually approximated with the output of a linear-feedback shift register with a known and very long period [9]). As a consequence, spreading sequences are defined as $\mathbf{y}_k^{m,n} = \mathbf{w}_u^{m,n} \mathbf{w}_u^{m,n} \mod N$.

Note that, since all codes corresponding to streams of the same user are scrambled by the same iid sequence they remain orthogonal over a period of $N$ chips. Yet, since the scrambling is iid, the overall spreading is also iid and this grants some immunity from asynchronous interference.

In the chaos-based approach the symbols $\mathbf{z}_u^{m,n}$ for $u = 0, \ldots, U - 1, m = 0, \ldots, M - 1$ and $k = 0, \ldots, N - 1$ are taken from the quantization of the trajectory of a discrete-time chaotic dynamical system. Formally speaking we will assume that a state-update function $M : [0, 1] \to [0, 1]$ exists.
The theoretical limits of asynchronous multi-code DS-CDMA systems over AWGN channels have been investi-
giving rise to the trajectory $x_{t+1} = M(x_t)$ starting from an initial condition $x_0$. For each $u, m$ pair, $x_0$ is drawn independently and the trajectory $x_0, \ldots, x_{N-1}$ is then quantized by the function $Q : [0, 1) \to \{-1, +1\}$ to obtain $z_{k,m} = Q(x_k)$.

Figure 2 shows a family of possible functions $M$ defined by a parameter $\lambda$. If $Q$ is such that $Q(x) = -1$ when $x \in [0, 1/2]$ and $Q(x) = +1$ when $x \in [1/2, 1]$ then it can be proved (see for example [10] and references therein) that $E[H_i H_j] = \lambda_i \delta_{ij}$ if $i < j$ and $E[H_i H_j] = \lambda_i \delta_{ij} - \lambda_j \delta_{ij}$ if $i \leq j \leq l$. These expressions for second- and fourth-order correlation can be generalized to any order thus leading to iid antipodal sequences when $\lambda = 0$.

In the following, we will analyze three kinds of sequence assignment for a multi-code DS-CDMA system with $U$ users and $M$ streams, all based on the availability of the $N$-symbols antipodal sequences $z_{k,m}^\alpha$ for $\lambda = -2 + \sqrt{3}$ that, in the case $M = 1$, are known to minimize the asynchronous interference between users [1]. The three methods are:

1. **C** We directly set $y_{k,m}^\alpha = z_{k,m}^\alpha$.
2. **CGS** For each $u = 0, \ldots, U - 1$ we apply the Gram-Schmidt orthonormalization procedure to the set of sequences $z_{k,m}^\alpha$ to obtain the set $z_{k,m}^{\alpha u}$. Then $y_{k,m}^\alpha = z_{k,m}^{\alpha u} \mod N$.

**CGS($q$)** The values of the orthonormalized sequences $z_{k,m}^{\alpha u}$ of CGS are first quantized with a finite number of bits $q$ to obtain $z_{k,m}^{\alpha u,q}$. Then $y_{k,m}^\alpha = z_{k,m}^{\alpha u,q} \mod N$.

The two last methods are inspired to what has been proposed in [11]. Note that though in principle CGS is unsuitable for implementation since the spreading sequences would features real-valued chips, its performance can be hopefully approximated by CGS($q$) with relatively small $q$. The case $q = 1$ is also extremely interesting since it yields antipodal spreading sequences though it is not clear in general whether the aim of mixing good counteracting of asynchronous interference with good orthogonality can be achieved with such a coarse re-quantization.

### 4. Simulation setup

We have found that the capacity is a function of the matrix $H$ that, in turn, depends on time-shifts $r^\alpha$ and spreading sequences $y_{k,m}^\alpha (u = 0, \ldots, U - 1, m = 0, \ldots, M - 1, k = 0, \ldots, N - 1)$, that, in the most general setting, are assumed to be $UM$ independent realizations of a discrete-time finite-value stochastic process with known statistics. With this, the capacity is itself a random variable and performance evaluation must take into account its statistics.

A C++ program has been built to simulate the system. It applies random delays and employs different quantized slices of chaotic trajectories for each run. After the capacity values are available, their statistics are highlighted by computing the CCDF (Complementary Cumulative Distribution Function) that is $Pr(C > C_0)$ where $C_0$ is a capacity value. Statistics are computed over 100,000 trials. After this, the mean capacity value is also computed.

### 5. Simulation results

Simulations have been conducted for $U = 13 \div 17$ and $M = 1 \div 5$ using OVSF+S, C, CGS, CGS(1) and CGS(5) spreading strategies.

We report in Table 2 the capacity values averaged over all the 100,000 trials. Note that chaos-based spreading is better than OVSF+S in all configurations.

Among chaos-based cases, note that CGS is better than C while CGS(5) is approximatively the same as CGS.

This highlights the importance of considering both asynchronous and synchronous aspects when the overall performance has to be improved and that 5 bits are enough to reproduce the behavior of continuous-valued sequences. Orthonormalization with 1-bit re-quantization offers the best result for $M < 5$. To highlight this strange behavior, Figure 3 reports a comparison between C, CGS and CGS(1) spreading strategies in terms of the percentual gain over OVSF+S plotted against $M$ for different values of $U$. Unexpectedly, the CGS(1) case seems to optimize both synchronous and asynchronous performance.

Figures 4 and 5 reports the resulting CCDF curves for the boundary cases of interest ($U, M = (13, 2), (17, 5)$). In all cases the curves of either CGS(1) or CGS(5) are higher than the OVSF+S curve. Hence, there is always a chaos based strategy completely outperforming OVSF+S. As an alternative quantification of the improvement we may use those curves to find that the probability of achieving a capacity larger than the average one for OVSF+S is approximately 50% for the OVSF+S itself while it is 94.8% and 86.8% for the best chaos-based method.

### 6. Conclusion

The theoretical limits of asynchronous multi-code DS-CDMA systems over AWGN channels have been investi-
Table 2: Capacity values averaged over 100,000 trials

<table>
<thead>
<tr>
<th>$U$</th>
<th>$M$</th>
<th>OVSF+S</th>
<th>C</th>
<th>CGS(1)</th>
<th>CGS(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1</td>
<td>66.2</td>
<td>126.2</td>
<td>180.8</td>
<td>230.3</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>66.5</td>
<td>127.5</td>
<td>183.1</td>
<td>233.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>66.5</td>
<td>127.8</td>
<td>184.0</td>
<td>235.4</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>66.5</td>
<td>133.0</td>
<td>194.8</td>
<td>243.3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>66.5</td>
<td>127.7</td>
<td>183.9</td>
<td>235.4</td>
</tr>
</tbody>
</table>

Figure 3: Comparison between C and CGS(1) spreading strategies in terms of the percentage of the mean capacity increment with respect to OVSF spreading as a function of $M$.

Figure 4: CCDF curves for $U = 13$ and $M = 2$.

Figure 5: CCDF curves for $U = 17$ and $M = 5$.

gated by computer simulations using the statistics of the associated Shannon capacity. Classical OVSF+S strategy has been compared with different chaos-based approaches. Results show that there is a chaos-based solution largely outperforming OVSF+S for each tested configuration. The unexpected good performance of CGS(1) will be the topic of further investigations.

References

An $M$-ary Orthogonal-CSK communication system

Pascal Chargé and Danièle Fournier-Prunaret

LESIA – INSA
135 avenue de Rangueil, 31077 Toulouse cedex 4, France
Email: Pascal.Charge@insa-toulouse.fr, Daniele.Fournier@insa-toulouse.fr

Abstract—In this paper we propose a secure multiple-access technique chaos-based digital communication system. In this system, a piecewise-linear map is used to generate a chaotic sequence, which is then projected onto two orthogonal basis functions. Data information are coded by switching between these orthogonal chaotic sequences. Only the sign bit of these chaotic signals is transmitted. Implementation details and security are discussed. Simulations of the noise performance of the system are also proposed.

1. Introduction

The use of chaotic signals for the transmission of data information has been of great interest in the past few years and many schemes have been proposed. Chaotic spread spectrum techniques [1] have been proposed. Chaos shifted-keying (CSK) systems and differential CSK (DCSK) systems have been reported [2][3]. It was shown in [4] that by combining the chaotic waveform with Walsh functions, a binary DCSK system may be modeled by means of two orthogonal basis functions, where each basis function represents a given bit. More recently, in [5], a subspace theory for DCSK systems has been proposed using orthonormal vectors to generate two orthogonal symbol subspaces corresponding to two different symbols.

In this paper we propose a multiple-access CSK-based communication system also using orthogonal Walsh functions. A chaotic map is used to generate the chaotic signal, which is then projected onto orthogonal Walsh functions during each symbol duration. Data information is coded by switching between these orthogonal chaotic sequences. Depending on the number $p$ of bits coded in every symbol, $M = 2^p$ Walsh functions will be required and the proposed system can then be qualified as an $M$-ary orthogonal chaotic shift keying (M-OCSK) system. The transmitted waveform is directly provided by the sign of the output sequence. The receiver is a coherent correlation demodulator. This M-OCSK system is a multiple-access communication system in that sense it allows the transmission of data by multiple users simultaneously.

The paper is organized as follows. The proposed M-OCSK system is firstly presented, then some elements of implementation and a brief security analysis are provided. In section 3 computer simulations are performed. Finally, section 4 concludes the paper.

2. System model

In this section, we present a multiple-access chaos-based technique under a noisy environment. Figure 1 shows the schematic baseband model of the system.

![Figure 1: Multiple-access system model](image)

2.1. Transmitted signals model

At each transmitter, the user must be able to send a binary information by switching between $M$ orthogonal chaotic sequences (with $M = 2^p$). Assume that a unique chaotic sequence $\{x_k\}_{k=0}^1$ is generated by a chaotic map. The $M$ orthogonal sequences will be obtained from that chaotic sequence. The $M$ orthogonal binary sequences are obtained from the sign bit elements of the sequences $\{w_k^i(0)\}$ where

$$\{w_k^i(0)\} = \{(-1)^{[2^{i-1}]}\}$$  \hspace{1cm} (1)

is the $i$th Walsh sequence with $0 \leq i \leq M - 1$. These $M$ distinct binary sequences can be used to represent $M$ different symbols (sets of $p$ bits).

The signal $s(t)$, transmitted by the $i$th user (see figure 1), is organized into symbol time slots which are sub-divided into chip time slots. $T_b$ is the symbol time slot duration, and $T_c$ is the chip time slot duration. The output of the chaotic map is changing every $T_c$ during the transmission. The choice of this chaotic map will be discussed later in this paper. A sign operator is used at the output of the chaotic signal generator, so that only the sign bit of the chaotic sequence is used.

The $m$th symbol to be transmitted during a period $T_b$ by the $i$th user is denoted by $d_{m,i}^i$ ($m = 0, 1, 2, 3, \ldots$). Let $\beta$ be the spreading factor of the system, defined as $T_b/T_c$, which is an integer. This spreading factor $\beta$ must be chosen such that $\beta = 2^\alpha$ with $\alpha \geq M - 1$, $\alpha$ being an integer.
Transmitted symbols \( d_{m}^{(i)} \) are letters of a finite alphabet of \( M \) elements. As an example if \( M = 4 \), each transmitted symbol is a set of \( p = 2 \) bits \( \{00, 01, 10, 11\} \). The \( l \)th symbol \( (0 \leq l \leq M - 1) \) of the finite alphabet is then represented by the binary signal \( s_{i}^{(l)}(t) \)

\[
s_{i}^{(l)}(t) = \sum_{m=0}^{\beta-1} \sum_{n=0}^{\infty} r(t - (m\beta + n)T_c)n_{w}^{(i)} \text{sign}(x_{n}^{(m,i)})
\]  
(2)

where \( x_{n}^{(m,i)} = (m\beta + n)^{th} \) sequence element of \( \{x_{k}\} \) in the transmitter \( i \) (the \( n^{th} \) chip during the bit period \( m \)), and \( r(t) \) is a rectangular pulse of unit amplitude and width \( T_c \).

According to properties of the Walsh sequences, it appears that \( \int_{T_c}^{T_c + T_{s}} s_{i}^{(l)}(t)x_{n}^{(m,i)}(t)dt = 0, \) \( \forall r \), when \( l \neq q \), and then the \( M \) binary signals are perfectly orthogonal with each other over a bit period. The overall transmitted waveform by the user \( i \), \( s^{(i)}(t) \), is

\[
s^{(i)}(t) = \sum_{m=0}^{\beta-1} \sum_{n=0}^{\infty} r(t - (m\beta + n)T_c)\text{sign}(x_{n}^{(m,i)})
\]  
(4)

where \( y_{n}^{(m,i)} = w_{n}^{(i)}x_{n}^{(m,i)} \) when \( d_{m}^{(i)} \) is the \( l^{th} \) symbol \( (0 \leq l \leq M - 1) \).

Assume that channel distortion is only due to a noise source \( n(t) \), which is an additive white Gaussian noise (AWGN) with a two-side power spectral density given by \( \gamma_{n}(f) = \frac{N_0}{2} \), \( \forall f \). The input signal at the receiver can be written as

\[
\hat{s}(t) = \sum_{i=1}^{N} s_{i}^{(i)}(t) + n(t)
\]  
(5)

where \( N \) is the number of users.

### 2.2. Pure coherent receiver

Receiver structure of the \( j^{th} \) user (see figure 1) is very close to the structure of a coherent CSK demodulator. There is only one chaotic generator and it is assumed that it can perfectly reproduce the chaotic signal of the \( j^{th} \) transmitter. It is also assumed that at the beginning of a transmission, the initial condition and the chaotic map parameters for the user \( j \) have been secretly transmitted by another way and are then only known at the receiver \( j \). Moreover, bit period synchronization is assumed. In this \( j^{th} \) receiver, the \( M \) orthogonal binary signals \( s_{i}^{(j)}(t) \) (equation (2)) are reproduced from the chaotic map.

Then correlations between the overall received signal at \( j^{th} \) receiver \( \hat{s}(t) \), and the \( M \) orthogonal signals \( s_{i}^{(j)}(t) \), are done during a bit period \( T_b \) in \( M \) separate correlators, and are given by

\[
\text{Corr}_{j}^{(i)}((m + 1)T_b) = \int_{mT_b}^{(m+1)T_b} \hat{s}(t)s_{i}^{(j)}(t)dt
\]

Correlators are reseted every bit period \( T_b \). Output of these correlators are compared and the estimated transmitted symbol \( d_{m}^{(j)} \) is the symbol \( \hat{l} \) \( (0 \leq \hat{l} \leq M - 1) \), which is given by

\[
\hat{l} = \arg \max \left( \text{Corr}_{j}^{(i)}((m + 1)T_b) \right)
\]  
(7)

### 2.3. Chaotic map

A noninvertible piecewise-linear map will be used in order to generate chaos at each transmitter. This map \( f : [-1, 1] \mapsto [-1, 1] \) is given by

\[
\begin{align*}
\begin{cases}
    z_k &= K|x_k| + \phi^{(i)} \quad [\text{mod } 1] \\
    x_{k+1} &= \text{sign}(x_k) \left( 2z_k - 1 \right)
\end{cases}
\end{align*}
\]  
(8)

and it depends on parameters \( K \) and \( \phi^{(i)} \). \( K \) is a fixed positive integer. For the following \( K = 2 \) will be arbitrary chosen. \( \phi^{(i)} \) is the parameter for the \( i^{th} \) transmitter and its associate receiver, with \( 0 \leq \phi^{(i)} < 1 \) and \( \phi^{(i)} \neq \phi^{(j)} \) for \( i \neq j \). The chosen map is represented in figure 2 when \( K = 2 \) and \( \phi^{(i)} = 0.2 \). Due to the evident symmetry property, the initial condition \( x_{0}^{(i)} \) will be chosen so that \( 0 < x_{0}^{(i)} < \frac{1}{2} \). In order to determine if this transformation can be used for multisuser transmission, we have first to control its sensitivity to initial condition and parameter. A way to measure this sensitivity is the Lyapunov exponent of a trajectory given by the transformation. The transformation is piecewise-linear with \( (2K + 2) \) regions, and the slope in each region is always \( 2K \), which is always greater than unity so that the Lyapunov exponent is \( L_K = \ln(2K) > 0 \). This last claim is true for any initial condition and any authorized value of \( \phi^{(i)} \).
2.4. Elements of implementation

As a great majority of chaos-based communication systems, the proposed M-OCSK system will suffer from many drawbacks. If chaotic generators and especially the considered chaotic map are analogical circuits, we can say that it is obviously impossible to have perfect copies of them available at the transmitter and receiver side, which is a very obstructing problem for us. Also the decoding of the information signal at the receiver side is based on a perfect knowledge of the initial condition and parameters of the chaotic map. If such initial condition and parameters are analogical values, even if transmitted at the receiver side, they will always be affected by an error. In order to avoid these problems, the proposed system can be implemented in a digital circuit. It can be implemented on DSP or FPGA technology. A digital implementation with fixed-point operators drains many advantages. Initial condition and parameters of the chaotic map can then be 32 bits words if a standard fixed-point arithmetic unit is used for the implementation. This digital system allows that perfect copies of the chaotic maps generating sequences are available at the transmitter and the receiver side. However the digital implementation of the chaotic map brings us to the fact that digital calculated sequences are periodic, which is not a real problem if the period is very large. This has to be studied and, as mentioned in [6], the use of extended non-linear feedback shift registers based on chaotic maps is a reasonable solution to produce sequences with long period.

2.5. Secret keys and security

In this scheme, the values of $\phi(i)$, the parameter of the chaotic map, and $x_0^{(i)}$, the initial condition, will form the symmetric secret key. Recalling that $0 \leq \phi(i) < 1$ and $0 < x_0^{(i)} < \frac{1}{2}$, if a 32 bits fixed-point arithmetic is used for this system and if $K = 2$, then the key will be a 61 bits long word (31 bits for $\phi(i)$ and 30 bits for $x_0^{(i)}$). Any of the $2^{61}$ available keys can be used equivalently since as previously mentioned, all of them will provide a chaotic sequence (the key has no effect on the Lyapunov exponent). It can be shown that for two keys with the slightest difference, the two binary generated chaotic sequences are totally different. The secret key for each user can then be chosen randomly. As we talk about secret key, we have to provide some elements about the security of the proposed system. Some rules have been suggested in [7] in order to evaluate how a cryptosystem is secure. The choice of the chaotic map is determinant for the security of the system. A noninvertible piecewise linear map has been chosen here. It can be verified that for any authorized key (parameter and initial condition) the probability of having +1 or −1 at the output of the chaotic generator is 1/2. The Perron-Frobenius operator [8] may help us to bring a proof of this property. The Perron-Frobenius operator allows to describe the time evolution of the probability density function $\rho_k(x)$ of a set of discrete-time orbits $\{x_k\}$. For a one-dimensional map $f$, when the derivative is not singular, let consider the Perron-Frobenius operator $P_F$:

$$P_F \rho_k(x) = \sum_{y \cdot f(y) = x} \frac{d(f(y))}{dy} \rho_k(y) \quad (9)$$

This operator maps the probability density function $\rho_k(x)$ to $\rho_{k+1}(x)$ such that

$$\rho_{k+1}(x) = P_F \rho_k(x) \quad (10)$$

Let now define $\eta_{a,b}(x)$ (with $b > a$) a rectangular pulse function of unit amplitude and width $b - a$:

$$\eta_{a,b}(x) = \begin{cases} 1 & \text{when } a < x < b \\ 0 & \text{elsewhere} \end{cases} \quad (11)$$

Let assume that the initial condition can be chosen randomly between 0 and $\frac{1}{2}$ with a uniform law, then one has $\rho_0(x) = K^g_{\eta} (\frac{x}{2})$. Directly applying (9) with the chaotic map (8), it comes that $\rho_1(x) = \frac{1}{2} \eta_{-1,1}(x)$, for any $K$. Using (9) again, it comes that when $k \geq 1$, $\rho_k(x) = \frac{1}{2} \eta_{-1,1}(x)$ whatever the parameter and the initial condition. Then the output binary sequence of the generator is statistically indistinguishable from the output of a truly random function with a uniform law, and is statistically the same for all keys. It comes also that partial knowledge of the key does not reveal any information about the transmitted data nor the unknown part of the key. Moreover, $\rho_k(x)$ being a uniform and an invariant probability density function (does not depend on $k$), it comes that if the sign of $x_k$ is known, $x_{k+1}$ can be positive or negative with the same probability for any key. So, binary sequences $\{w_{i}^{(1)}x_k\}$ and $\{w_{i}^{(2)}x_k\}$ (with $i_1 \neq i_2$) have the same statistical property for all keys. The switching modulation between these two sequences is then undetectable with basic spectral or statistical analysis, and the transmitted symbols can not be detected easily.

In order to claim that the proposed system is really secure, it should be checked whether it can be broken by all known attacks, which is not provided in this paper. According to that, and since the key size space is $2^{61}$ (not sufficiently secure against brute-force attacks [7]), we are not able to claim that the proposed system is a very strongly secure system. It can nevertheless be used when a low security is required such as in many embedded wireless communication systems.

3. Computer simulations

Some computer simulations are now performed. For these simulations, parameter and initial condition for the chaotic map of each user are randomly generated, $K = 2$, and 32 bits fixed-point arithmetic units are used to generate the chaotic sequences. In figure 3, BERs are calculated from $10^6$ transmitted symbols for the M-OCSK.
with \( M = 2 \) and 4 in the multiuser case. Performance of the M-OSCK in the single user case is also compared with the theoretical bounds provided by the conventional coherent antipodal BPSK system [2]. Noise performance of the M-OCSK in the 10 users is shown in figure 4 when \( \beta \) is varying.

![Figure 3: BER of M-OCSK with M=2 and M=4 for different values of N the number of users.](image)

![Figure 4: BER of M-OCSK with different values for \( \beta \)](image)

It can be shown that the DCSK system (also using Walsh functions) presented in [4], and the M-OCSK system when \( M = 2 \) in the single user case, show exactly the same performance.

4. Conclusion

In this paper a so-called M-OCSK multiuser chaos-based communication system has been presented. In this system, only one chaotic generator is used to provide the \( M \) switching sequences, which is a substantial advantage in terms of reduction of the complexity. Moreover, the fact that the switching sequences are orthogonal contributes to the improvement of the performance compared with that of a conventional CSK system [2]. The M-OCSK system can be used concurrently to a standard DS-CDMA system or even jointly to a standard DS-CDMA system without bringing any perturbation, excepted obviously multiuser interferences. Moreover, the M-OCSK multiuser system has the advantage over a standard DS-CDMA that it can be presented as a secure communication system.

Acknowledgment This work was supported by the French project ACI-SI (Action Concertée Incitative - Sécurité Informatique) TRANSCHAOS no. 4E01.

References


A connection between chaotic message-embedding and conventional self-synchronizing stream ciphers

G. Millérioux†, J.M Amigó‡ and J. Daafouz§

†Université Henri Poincaré, France
Centre de Recherche en Automatique de Nancy (CRAN UMR CNRS 7039)
‡Universidad Miguel Hernández, Spain
Centro de Investigación Operativa
§Institut National Polytechnique de Lorraine
Centre de Recherche en Automatique de Nancy (CRAN UMR CNRS 7039)
Emails: gilles.millerioux@estin.uhp-nancy.fr, jm.amigo@umh.es, jamal.daafouz@ensem.inpl-nancy.fr

Abstract—A lot of encryption methods involving chaotic dynamics have been proposed in the literature since the 90’s. Most of them consists of “mixing” the information to be hidden with a chaotic sequence. In this paper, a connection between one of the most attractive chaotic schemes, namely, hybrid message-embedding and the conventional self-synchronizing stream cipher is carried out. The main point can be stated as follows: hybrid message-embedding is strictly equivalent to a conventional self-synchronizing stream cipher under flatness conditions.

1. Introduction

Modern cryptography originates in the works of Feistel at IBM during the late 1960s and early 1970s. One of the key dates is 1977, when the symmetric (or private-key) algorithm called Data Encryption Standard (DES) was adopted by the U.S. National Bureau of Standards (now the National Institute of Standards and Technology —NIST), for encrypting unclassified information. DES is now in the process of being replaced by the Advanced Encryption Standard (AES), a new standard adopted by NIST in 2001. Another milestone is 1978, marked by the publication of the Flatness conditions.

$$\lim_{k \to \infty} \|T x_k - \hat{x}_k\| = 0 \ \forall x_0 \in U$$

or

$$\exists k_f < \infty, \ |x_k - \hat{x}_k| = 0 \ \forall x_0 \in U \ and \ k \geq k_f$$

where $T$ is a constant matrix of appropriate dimension and $U$ is a non empty set of initial conditions. (1) corresponds to an asymptotic synchronisation, while (2) corresponds to a finite time synchronisation. Let us point out that in practice, since we deal with finite accuracy, the error of an asymptotical synchronisation can be considered to be zero after a finite transient time.

2. Hybrid Message-embedding

The hybrid message-embedded technique (Fig. 1) was proposed in [14] and partially cryptanalyzed in [12] wherein the term “hybrid” was first introduced. We distinguish two distinct setups. The first one is governed by the state equations:

$$x_{k+1} = f_0(x_k, u_k)$$
$$y_k = h_0(x_k, u_k)$$
$$u_k = v_r(x_k, m_k)$$

while the second class is governed by:

$$x_{k+1} = f_0(x_k, u_k)$$
$$y_k = h'_0(x_k)$$
$$u_k = v_r(x_k, m_k)$$
The systems (3) and (4) differ from each other by their relative degree.

Definition 1 ([4] P.139) The relative degree of a system with respect to the quantity \( u \) is the required number \( r \) of iterations of the output \( y_k \), so as \( y_k \) depends on \( u \), which actually appears explicitly in the expression of \( y_{k+r} \).

Based on Definition 1, the relative degree of the system (3) is \( r = 0 \). The system (4) has a relative degree \( r \) strictly greater than 0. It means that, after iterating \( r \) times the state vector \( x_k \), the output \( y_{k+r} \) reads

\[
y_{k+r} = h_0(f_0^r(x_k, u_k))
\]

where

\[
f_0^r(x_k, u_k) = \begin{cases} x_k & \text{when } i = 0 \\ f_0(f_0^{i-1}(x_k, u_k), u_{k+i-1}) & \forall i \geq 1. \end{cases}
\]

and where \( u_k \) appears explicitly in the sense that there exists \( u'_k \neq u_k \) such that \( y_{k+r} = h_0(f_0^r(x_k, u_k)) \neq h_0(f_0^r(x_k, u'_k)) \) whereas for all \( u'_k \neq u_k, \ y_{k+r'} = h_0(f_0^r(x_k, u_k)) = h_0(f_0^r(x_k, u'_k)) \) if \( r' < r \). Let us point out that \( u_k \) is sometimes called the "pre-ciphertext".

The receiver system must be designed in such a way that both \( u_k \) and \( x_k \) can be recovered, given the only available data \( y_k \) and its subsequent iterates. Once \( u_k \) is recovered, the plaintext \( m_k \) is correctly extracted by applying the decryption function \( y_d \), provided that \( \hat{x}_k \) is exactly synchronized with \( x_k \). The synchronization and the recovering of \( u_k \) can resort to an inverse system or to an unknown input observer of the form

\[
\begin{align*}
\hat{x}_{k+1} &= \hat{f}_0(\hat{x}_k, y_1, \ldots, y_{k+r}) \\
\hat{u}_k &= g_0(\hat{x}_k, y_{k+r}) \\
\hat{m}_k &= y_d(\hat{x}_k, \hat{u}_k)
\end{align*}
\]

with \( g \) such that

\[
\hat{u}_k = g_0(\hat{x}_k, y_{k+r}) = u_k \quad \text{when} \quad \hat{x}_k = x_k
\]

and with \( y_d \) such that

\[
\hat{m}_k = y_d(\hat{x}_k, \hat{u}_k) = m_k \quad \text{when} \quad \hat{x}_k = x_k \quad \text{and} \quad \hat{u}_k = u_k.
\]

Unlike other classical methods, the hybrid message-embedded technique offers the advantages that only a single channel is needed and, moreover, that the synchronization can be guaranteed without restriction on the rate of variation of \( m_k \). Additionally, the scheme allows to introduce a highly nonlinear function \( y_d \), which can make the state generator significantly resistant to algebraic attacks.

![Diagram](image)
identical at both sides, the respective keystreams synchronize automatically because $\sigma^0$ operate, at both sides, on the same quantities, namely the past values of $c_i$. The ability to self-synchronize constitutes one of the main advantages of such cryptosystems. Indeed, they are resistant against bit slips on the transmission channel without any additional synchronization flags or interactive protocols for recovering lost synchronization.

3.2. Main result

First of all, we must recall the definition of flatness (see [8] for an introductory theory)

**Definition 2** (Flatness) A system with dynamic $f$, parametrized by relative degree $r$, input $e_k$ and state vector $z_k$ of dimension $n$ is said to be flat if there exists a set of independent variables $y_k$, referred to as flat outputs, such that all system variables can be expressed as a function of the flat output and a finite number of its backward and/or forward iterates.

For Single Input Single Output systems, it means that there exist two functions $F_\theta$ and $G_\theta$ which obey

$$\begin{cases}
z_k &= F_\theta(y_k, e_k, \ldots, y_k, e_{k_0}) \\
e_k &= G_\theta(y_k, e_k, \ldots, y_k, e_{k_0})
\end{cases} \tag{10}$$

where $k_F(r), k_E(r), k_G(r)$ and $k_{G'}(r)$ are $\mathbb{Z}$-valued integers depending on the relative degree $r$ of the system.

**Proposition 1** The hybrid message-embedding cryptosystem (3) (or (4)) is equivalent to a conventional self-synchronizing stream cipher if the nonlinear dynamic $f$ with output $y_k$ and input $u_k$ is flat.

**Proof 1** Flatness of (3), with relative degree $r = 0$, means that there exist two functions $F_\theta$ and $G_\theta$ and integers $k_F(0), k_E(0)$ and $k_{G'}(0)$ such that

$$\begin{cases}
x_k &= F_\theta(y_k, e_k, \ldots, y_k, e_{k_0}) \\
u_k &= G_\theta(y_k, e_k, \ldots, y_k, e_{k_0})
\end{cases} \tag{11}$$

When iterating once forward the first equation of (11), it turns out that (3) is strictly equivalent to

$$\begin{cases}
x_{k+1} &= F_\theta(y_k, e_k, \ldots, y_k, e_{k+1}) \\
y_k &= G_\theta(y_k, e_k, \ldots, y_k, e_{k_0})
\end{cases} \tag{12}$$

Letting $l_{h_{v_e}}(x_k, m_k) = h_{v_e}(x_k, v_e(x_k, m_k))$ since $y_k$ depends explicitly on $x_k$ and $m_k$, identification of (12) with (9) leads then to the following result:

i) The system (3) is strictly equivalent to the transmitter part of a self-synchronizing stream cipher of the form (9) with key generator $\sigma^0 = F_\theta$, running key $K_0 = x_k$, ciphertext $c_k = y_k$, encrypting function $e = l_{h_{v_e}}$, secret static key $\theta$ and memory $b = [k_{y_F}(0) - k_{y_F}'(0) + 1]$.

Besides, flatness of (4), with relative degree $r > 0$, means that there exist two functions $F_\theta$ and $G_\theta$ and integers $k_F(r), k_E(r)$ and $k_{G'}(r)$ such that

$$\begin{cases}
x_k &= F_\theta(y_k, e_k, \ldots, y_k, e_{k_0}) \\
u_k &= G_\theta(y_k, e_k, \ldots, y_k, e_{k_0})
\end{cases} \tag{13}$$

When iterating once forward the first equation of (13) and taking into account (5), it turns out that (4) is strictly equivalent to:

$$\begin{cases}
x_{k+1} &= F_\theta(y_k, e_k, \ldots, y_k, e_{k+1}) \\
y_k &= G_\theta(y_k, e_k, \ldots, y_k, e_{k_0})
\end{cases} \tag{14}$$

Letting $l_{h_{v_e}}(x_k, m_k) = h_{v_e}(f_{y_F}(x_k, v_e(x_k, m_k)))$ since $y_{k+r}$ depends explicitly on $x_k$ and $m_k$, identification of (14) with (9) leads then to the following result:

ii) The system (4) is strictly equivalent to the transmitter part of a self-synchronizing stream cipher of the form (9) with key generator $\sigma^0 = F_\theta$, running key $K_0 = x_k$, ciphertext $c_k = y_k$, encrypting function $e = l_{h_{v_e}}$, secret static key $\theta$ and memory $b = [k_{y_F}(r) - k_{y_F}'(r) + 1]$.

**Remark 1** It is worthwhile noticing that the set of equations (11) (resp. (13)) could be used at the receiver part to obtain both $x_k$ and $u_k$ without resorting to a state reconstruction through an inverse system or an Unknown Input Observer like (6). Even more is true: substituting $x_k$ and $u_k$ of (11) (resp. (13)) into (8) gives

$$m_k = v_e(F_\theta(y_k, e_k, \ldots, y_k, e_{k_0}), G_\theta(y_k, e_k, \ldots, y_k, e_{k_0})). \tag{15}$$

with $r = 0$ (resp. $r > 0$). So, the message $m_k$ can be retrieved in finite time by and the knowledge of $x_k$ is no longer useful. However, given a system, the difficulty lies in finding out the quantities $k_F(r), k_E(r), k_{G'}(r)$ and $k_{G'}(r)$ and writing down explicitly the functions $F_\theta$ and $G_\theta$. It can be shown (see [11] for the linear case) that resorting to a state space approach actually allows to achieve this computation in an implicit and recursive way. Indeed, for flat systems, only a finite number of iterations of (6) is needed to achieve $\hat{x}_k = x_k$. It turns out that the resulting state vector $\hat{x}_k = x_k$ only depends on past values of $y_k$, which provides $F_\theta$ in (10). Then, substituting $\hat{x}_k = x_k$ into (7) provides $G_\theta$ in (10).

4. Example

We consider a 3-dimensional linear congruential hybrid message-embedded cryptosystem like (4) with dynamic $f$ and output function $h'$ of the form:

$$\begin{cases}
x_{k+1} &= A x_k + B u_k \\
y_k &= C x_k \\
u_k &= v_{e}(x_k, m_k)
\end{cases} \tag{16}$$
The entries of the matrices $A$, $B$ and $C$ are integers ranging between 0 and 255, the modulo being $m = 256$. Along this section, the operations are performed modulo $m$.

Numerically, the matrices read

$$
A = \begin{bmatrix}
38 & 1 & 0 \\
7 & 0 & 1 \\
4 & 0 & 0
\end{bmatrix}, 
B = \begin{bmatrix}
1 \\
0 \\
0
\end{bmatrix}, 
C = \begin{bmatrix}
1 & 0 & 0
\end{bmatrix}.
$$

It is recalled that, for linear systems written in a state space form, the relative degree corresponds to the smallest integer $r$ such that $CA^{r-1}B$ is different from 0 ([4]). Here, since $CB = 1$, the relative degree of the system is 1. The supposed secret static key is the vector $\theta = [38 \ 7 \ 4]$ which actually corresponds to the first column of $A$ written in a companion form. The function $v_r$ is chosen to be a bitwise XOR (denoted $\oplus$) between the components of $x_k$ denoted $x_k^{(i)}$ and the plaintext $m_k$:

$$
u_k = x_k^{(1)} \oplus x_k^{(2)} \oplus x_k^{(3)} \oplus m_k.
$$

where $x_k^{(i)}$ and $m_k$ are meant here to be the corresponding 8-bit representation. It turns out that after iterating three times the inverse system of (16) (the structure is not provided here but see for example [1] for details), as mentioned in the Remark 1, we obtain the equations in the form (13) with $F_\theta$ obeying

$$
\begin{align*}
\begin{bmatrix}
x_k^{(1)} \\
x_k^{(2)} \\
x_k^{(3)}
\end{bmatrix} &= 
\begin{bmatrix}
y_k \\
7y_{k-1} + 4y_{k-2} \\
y_{k-1}
\end{bmatrix} \\
(17)
\end{align*}
$$

and the function $G_\theta$ obeying

$$u_k = y_{k+1} = 38y_k - 7y_{k-1} - 4y_{k-2} \quad . \quad (18)
$$

Equations (17) and (18) clearly corroborate that the system is flat. Besides, they provide the actual values $k_{F_\theta} = 0$, $k_{G_\theta} = 0$, $k_{F_\theta}(1) = -2$, $k_{G_\theta}(1) = 1$ and $k_{G_\theta}(2) = -2$. The relative degree $r$ of the system being 1, we must compute $y_{k+1}$:

$$
y_{k+1} = CAx_k + CBv_r(x_k, m_k) = I_{x_k} \oplus \nu_r(x_k, m_k) = 38x_k^{(1)} + x_k^{(2)} + x_k^{(1)} \oplus x_k^{(2)} \oplus x_k^{(3)} \oplus m_k \quad (19)
$$

Iteration of (17) once forward and consideration of (19) allow us to claim the result ii):

The system (16) is strictly equivalent to the transmitter part of a self-synchronizing stream cipher formed by the form (9) with key generator $\sigma_\theta^+: = F_\theta$ corresponding to Eq. (17), running key $K_k = x_k$, ciphertext $c_k = y_{k+1}$, encrypting function $\nu = I_{y_r} \oplus I_{y_r}$, corresponding to Eq. (19), secret static key $\theta = [38 \ 7 \ 4]$ and memory $b = 2 + 1 = 3$.

Retrieving $m_k$ requires the computation (15). Here the function $v_d$ is also an XOR between the components of $x_k$ and the (pre-)ciphertext $u_k = v_d(x_k, m_k)$, that is, $v_d(x_k, u_k) = u_k \oplus x_k^{(1)} \oplus x_k^{(2)} \oplus x_k^{(3)}$. Indeed, $u_k \oplus x_k^{(1)} \oplus x_k^{(2)} \oplus x_k^{(3)} = m_k \oplus x_k^{(1)} \oplus x_k^{(2)} \oplus x_k^{(3)} = m_k$. The system being flat, $x_k^{(i)}$ can be expressed in terms of delayed outputs as indicated by the function $F_\theta$. Hence, one has

$$
m_k = (y_{k+1} - 38y_k - 7y_{k-1} - 4y_{k-2}) \oplus y_2(7y_{k-1} + 4y_{k-2}) \oplus 4y_{k-1}.
$$

5. Conclusion

We conclude, based on the parallelism mentioned above, that digital hybrid message-embedding is able to provide the same security as any conventional self-synchronizing stream cipher, currently being used (e.g., RC4) in, say, internet and mobile communications, under some suitable choice of functions $f$, $h$ (or $h'$) and $v_r$.

References


Lossless Parallel Image Coding using Discrete Time Cellular Neural Networks

Ryunosuke Okamura*, Hisashi Aomori*, Tsuyoshi Otake†, Nobuaki Takahashi‡ and Mamoru Tanaka*

*Department of Electrical and Electronics Engineering, Sophia University, 7-1, Kioi-cho, Chiyoda-ku, Tokyo 102–8554, Japan
Email: ryunosuke@tlab.ee.sophia.ac.jp
†Department of Media-Network Science, Tamagawa University, 6-1-1, Tamagawagakuen, Machida, Tokyo, 194–8610, Japan
‡IBM Engineering & Technology Services, IBM Japan, Ltd., 1623-14, Shimotsuruma, Yamato-shi, Kanagawa 242–8502, Japan

Abstract — The lifting scheme is a flexible method for the construction of linear and nonlinear wavelet transforms. In the nonlinear lifting scheme, it is difficult to design the optimal update filter corresponding to the nonlinear prediction filter. In previous work, we proposed a lossless coding method using Discrete Time Cellular Neural Network (DT-CNN). However, this method takes a lot of processing cost. In order to reduce the processing time, we propose a novel parallel image coding using DT-CNN. In our proposed method, we divide an image into some sub-images, and consider each sub-image to be independent one small image. To implement the processing of some sub-images in parallel, we can solve the bottleneck that the image processing cost is high. The experimental results show that the processing cost is greatly reduced by the proposed coding scheme. Our proposed method enables to implement hardware with high performance by reducing the cost of memories.

1. Introduction

The lifting scheme [1] is a general framework for constructing biorthogonal wavelets, and it has been recognized that nonlinear extensions are possible [2]. The main features of the lifting scheme are that it provides a spatial domain interpolation of the transform entirely and it can be extended into the hierarchical structure easily. Since it also provides reversible wavelet transforms for lossless image and signal compression, it has been applied to many applications such as remote sensing and medical imaging. The performance of the lifting method depends on the ability of the filters to interpolate images. In the conventional lossless image coding using the lifting method, the degradations are caused by the use of the integer wavelet transform instead of the discrete wavelet transform [3]. For efficient interpolations, the quantization noises propagated by the rounding operations should be considered.

Discrete-time cellular neural networks (DT-CNNs) [4] have been applied to many applications such as image compression, filtering, and pattern recognition [5], [6]. The nonlinear interpolative dynamics by feedback A-template is one of the significant characteristics of CNN. In some cases, however, the interpolative dynamics of DT-CNN is not used effectively because the model works as a linear filter. In [5], the nonlinear interpolative dynamics by feedback A-template of DT-CNN was used for image compression and the image interpolation corresponded to the optimization problem minimizing the Lyapunov energy function. In other words, the DT-CNN is a solver to solve the optimization problem to minimize the Lyapunov energy function.

In our previous work [6], we proposed a lossless coding method using DT-CNNs. Although this method had a good lossless coding performance, the image processing cost was high. To solve the weak point, we propose a parallel coding method. Our proposed method is to divide an image into some sub-images and to implement each sub-image processing in parallel. By implementing some sub-image in parallel, the image processing times are greatly reduced. Our proposed method enables to implement hardware with high performance by reducing the cost of memories. The experimental results shows that our proposed method has a good lossless coding efficiency and reduce the image processing cost.

2. Discrete-Time Cellular Neural Network

Fig. 1 shows the block diagram of the DT-CNN. The state equation of DT-CNN is described in matrix form as

\[ x_{n+1} = Af(x_n) + Bu + T, \]  
\[ y_{n+1} = f(x_{n+1}), \]  

(1)  
(2)
where \( u \) is the input matrix, \( x \) is the state variable, \( f() \) is the multi-level quantizing function, and \( T \) is the constant matrix. \( A \) and \( B \) are feed-back and feed-forward template matrices, respectively. The Lyapunov energy function of DT-CNN [5] is defined by

\[
E_t = -\frac{1}{2} y^T (A - \delta I) y - y^T B u - T^T y. \tag{3}
\]

where \( \delta \) is the positive constant value to determine the quantizing region such that \( x = \pm \delta \) for \( f(x) = \pm 1 \). It is proved that the Lyapunov energy function becomes monotonic decreasing function, if the \( A \) matrix is symmetric and the diagonal elements are larger than zero [5].

In order to obtain the high quality image, it is necessary that image can be reconstructed considering the distortion. Let \( G \) be a Gaussian filter, the distortion function is defined by

\[
dist(y, u) = \left\| \frac{1}{2} y^T (G y - u) \right\|. \tag{4}
\]

It means that the difference between the interpolative prediction image and the input image should be small.

3. Lossless Parallel Image Coding using DT-CNNs

Fig. 2 shows the image dividing method of our proposed system. The input image \( u \) is divided into some sub images. As each divided sub image can be implemented independently, we consider each sub image to be a input of the encoder using DT-CNN.

Fig. 3 shows the block diagram of our system. At the split stage, the input image \( u_o \) is divided into even polyphase components \( u_{on} \) and odd polyphase components \( u_{en} \) such like Fig. 4. In this case, in consideration of the influence of the quantization error propagated by the rounding operations, the horizontal interpolation is applied to only the divided even polyphase components. The prediction for each \( u_{on} \) is designed by using the two-layered DT-CNN. In the first layer DT-CNN, the distortion function is minimized, and the compressed image for reconstruction can be obtained by the nonlinear interpolative effect of A-template. In the second layer DT-CNN, the interpolated image is obtained by DT-CNN filtering using B-template. Then we obtain the prediction residual \( c_n \) transmitted to the decoder. The update for each \( u_{en} \) is designed by using the DT-CNN which has no dynamics. Then we obtain the updated image \( c_n \) which is the input of the next stage. In the encoder, these lifting processes using the DT-CNN are applied to the even polyphase image iteratively. In the decoder, the same lifting rules are applied, and the reconstruction image is gradually improved by adding the difference image and the interpolated components.

3.1. Image interpolation using two-layered DT-CNN

The subsampled even polyphase images such like Fig. 4 are interpolated using the two-layered DT-CNN. By the comparison between (3) and (4), the templates and the parameters of the first layer DT-CNN are determined as

\[
A = A(i, j; k, l), \quad C(k, l) \in N_r(i, j) \tag{5}
\]

\[
\begin{cases} 
-(1 + \lambda) & \text{if } k = i \text{ and } l = j, \\
1 - \frac{1}{2\pi\sigma^2} \exp \left( -\frac{(l - i)^2 + (k - j)^2}{2\sigma^2} \right) & \text{otherwise},
\end{cases}
\]

\[
B = B(i, j; k, l), \quad C(k, l) \in N_r(i, j) \tag{6}
\]

\[
\begin{cases} 
1 & \text{if } k = i \text{ and } l = j , \\
0 & \text{otherwise}.
\end{cases}
\]
3.2. Image Update using linear filter

In order to avoid aliasing, we use the Le Gall 5-tap linear filter for the update steps. The updated image \( c_{ij} \) is obtained by

\[
    c_{ij} = \begin{cases} 
    u_{e_{ij}} + [(c_{i-1,j} + c_{i,j} + 2)/4] & \text{odd layer stages}, \\
    u_{e_{ij}} + [(c_{i-1,j} + c_{i,j} + 2)/4] & \text{even layer stages}, 
    \end{cases}
\]

(11)

where \([·] \) denotes the round-off operator.

4. Experimental Results

We implemented the coder and decoder of our proposed lossless image coding algorithm. We applied our system to the 8-bit standard gray-scale test images: “man,” “boat,” “elaine,” “lx,” “lena,” and “tiffany.” The size of the image “man” is 1024×1024 pixels, the others are 512×512 pixels.

The performance of the proposed method was compared with the separable 2D lifting method using Le Gall 5-tap/3-tap filters (JPEG 2000) which is indicated as “5/3”, S+P transform which is represented as “S+P”, and our previous method [6]. For the simulation, the coding factors are decided experimentally; the number of lifting layers \( L = 12 \), the \( r \)-neighborhood of cell \( r = 2 \), the regularization parameter \( \lambda = -1 \), and Table 1 shows each of the standard deviation of Gaussian.

Table 2 shows the entropy of each image coding. We can confirm that our proposed method almostly outperforms those of the conventional lifting methods.

We compared the entropy and the processing time of our proposed parallel image coding method with the method the previous method. The form of a image dividing at tiling is \( 2 \times 2 \). Table 3 shows that our proposed method reduce the processing time extremely compared with our previous method, and keep nearly high performance of our previous method. It means that the processing cost is greatly reduced by the proposed method.

5. Conclusion

The lossless parallel image coding using DT-CNNs has been proposed. In our previous method, lossless coding method using DT-CNNs had a good lossless coding performance compared with that of conventional lifting method. However, the image processing...
Table 2: Entropy of each coded image (bit/pixel).

<table>
<thead>
<tr>
<th>Image</th>
<th>Original</th>
<th>5/3</th>
<th>S+P</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>7.52</td>
<td>4.86</td>
<td>4.85</td>
<td>4.85</td>
</tr>
<tr>
<td>boat</td>
<td>7.19</td>
<td>5.02</td>
<td>5.08</td>
<td>5.00</td>
</tr>
<tr>
<td>elaine</td>
<td>7.51</td>
<td>5.00</td>
<td>5.13</td>
<td>4.96</td>
</tr>
<tr>
<td>lax</td>
<td>6.83</td>
<td>5.89</td>
<td>5.96</td>
<td>5.90</td>
</tr>
<tr>
<td>lena</td>
<td>7.45</td>
<td>4.42</td>
<td>4.40</td>
<td>4.36</td>
</tr>
<tr>
<td>tiffany</td>
<td>6.60</td>
<td>4.34</td>
<td>4.36</td>
<td>4.34</td>
</tr>
</tbody>
</table>

Table 3: Comparison of the entropy and the processing times between proposed method and previous method.

<table>
<thead>
<tr>
<th>Image</th>
<th>Method</th>
<th>Entropy</th>
<th>Processing time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>man</td>
<td>previous</td>
<td>4.85</td>
<td>24.27</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>4.85</td>
<td>5.73</td>
</tr>
<tr>
<td>boat</td>
<td>previous</td>
<td>5.00</td>
<td>4.78</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>5.00</td>
<td>1.28</td>
</tr>
<tr>
<td>elaine</td>
<td>previous</td>
<td>4.96</td>
<td>5.05</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>4.96</td>
<td>1.30</td>
</tr>
<tr>
<td>lax</td>
<td>previous</td>
<td>5.90</td>
<td>4.14</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>5.90</td>
<td>1.17</td>
</tr>
<tr>
<td>lena</td>
<td>previous</td>
<td>4.45</td>
<td>5.39</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>4.46</td>
<td>1.46</td>
</tr>
<tr>
<td>tiffany</td>
<td>previous</td>
<td>4.34</td>
<td>5.20</td>
</tr>
<tr>
<td></td>
<td>proposed</td>
<td>4.34</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Acknowledgment

This research is supported by the fund of Open Research Center Project from MEXT of Japanese Government (2002–2006).

References


Image Resolution Enhancement using Discrete-Time Cellular Neural Network

Takefumi Konishi*, Hisashi Aomori**, Tsuyoshi Otake*, Nobuaki Takahashi†, Seiya Masuda* and Mamoru Tanaka**

*Department of Media-Network Science, Tamagawa University,
6-1-1, Tamagawagakuen, Machida, Tokyo, 194-8610 Japan
**Department of Electrical and Electronics Engineering, Sophia University,
7-1, Kioi-cho, Chiyoda-ku, Tokyo, 102-8554 Japan
†IBM Engineering & Technology Services, IBM Japan, Ltd.,
1623-14, Shimotsuruma, Yamato-shi, Kanagawa, 242-8502 Japan
Email: knsta1ec@engs.tamagawa.ac.jp

Abstract—In this paper, novel image resolution enhancement technique using discrete-time cellular neural network (DT-CNN) is proposed. In our proposed method, images are interpolated by exploiting the nonlinear interpolative dynamics of the DT-CNN with a feedback A-template effectively. First, the DT-CNN transforms all pixel values into the optimal coefficients which make possible to establish the optimal prediction of the original pixel values using the A-template. Then, the optimal interpolated image is obtained by using the convolution of the B-template which is derived by extending the A-template spatially. The experimental evaluation shows that the proposed method produces better results than the conventional image resolution enhancement methods.

1. Introduction

Image resolution enhancement and interpolation are used to generate a high-resolution (HR) image from its low-resolution (LR) version. It is indispensable for high-resolution image processing applications such as digital HDTV, digital photogrammetry, high-quality printing, medical imaging, military purpose imaging, and so on. Recently wavelet based interpolation methods are discussed in many literatures [1]-[4]. In wavelet based techniques, it is assumed that the LR image to be resolution enhanced is the lowpass-filtered and downsampled HR image. In other words, the LR image is corresponding to the low-frequency coefficients of a wavelet transformed HR image. These methods are based on wavelet-domain zero padding (WZP) [4]. In the WZP, the LR image values are set to low-frequency subband of wavelet transform and the high-frequency subbands composed of all-zero. Therefore, due to the constraint of the dyadic decomposition for wavelet transform, its resolution enhancement factor is limited to 2^n, (n=1, 2, ...).

In this paper, we propose an image resolution enhancement technique with arbitrary enhancement parameters using a discrete-time cellular neural network (DT-CNN). The DT-CNN has been applied to many applications such as image compression, filtering and recognition [5]-[11]. The nonlinear interpolative dynamics by feedback A-template is one of the significant characteristics of CNN, it is a solver to solve the optimization problem to minimize the Lyapunov energy function. In our proposed method, the interpolation is based on the dynamics of the DT-CNN, where the output function is utilized in order to consider the nonlinear quantization error for optimization. At the equilibrium state of DT-CNN, we can obtain the optimal coefficients to be established an optimal prediction image.

2. Image Resolution Enhancement using DT-CNN

2.1. DT-CNN

Figure 1 shows the block diagram of DT-CNN. The state equation of the DT-CNN is described in matrix form as

$$x_{n+1} = Af(x_n) + Bu + T,$$

where $u$ is an input vector, $x$ is a state variable, $f()$ is a multi-level quantizing function, $A$ and $B$ are feedback and feedforward template matrices, respectively, and $T$ is a constant vector. Let $y=f(x)$, then the Lyapunov energy function $E$ of the DT-CNN is defined by

$$E = -\frac{1}{2}y'(A-\delta I)y - y'Bu - Ty'y,$$

where $\delta$ is a positive constant value to determine the quantizing region. If the $A$ matrix is symmetric and its diagonal elements are larger then zero, it is proved that the Lyapunov energy function becomes a monotone decreasing function [12] [13].

Figure 1: Discrete-time cellular neural network.
2.2. Image Resolution Enhancement using Two-Layered DT-CNN

In the proposed method, images are interpolated using the two-layered DT-CNN. At the first layer DT-CNN, in order to obtain the high accuracy prediction images, it is necessary that the image can be reconstructed based on the distortion defined by

$$\text{dist}(y,u) = \frac{1}{2} y' (Gy - u),$$  \hspace{1cm} (3)

where $G$ is a Gaussian filter. This distortion means that the difference between the interpolative predicted image and the input image should be small. By the comparison between equation (2) and (3), the A-template, B-template, and the constant $T$ can be determined as

$$A(i,j,k,l), \quad C(k,l) \in N_r(i,j)$$

$$= \begin{cases} 
(1 + \lambda), & \text{if } k = i \text{ and } l = j, \\
- \frac{1}{2 \sigma^2 \pi} \exp \left( -\frac{((k-i)^2 + (l-j)^2)}{2 \sigma^2} \right), & \text{otherwise},
\end{cases}$$  \hspace{1cm} (4)

$$B(i,j,k,l), \quad C(k,l) \in N_r(i,j)$$

$$= \begin{cases} 
1, & \text{if } k = i \text{ and } l = j, \\
0, & \text{otherwise},
\end{cases}$$  \hspace{1cm} (5)

$$T = 0,$$  \hspace{1cm} (6)

where $\sigma$ is the standard deviation of Gaussian function and $\lambda$ is a regularization parameter. The B-template is only nonzero at the center value. Then we can represent the dynamics of the first layer DT-CNN by using the above parameters as follows:

$$x_{ij} (t + 1) = \sum_{(k,l) \in N_r(i,j)} A(i,j,k,l) y_{kl}(t) + u_{ij},$$

$$y_{ij} (t + 1) = f(x_{ij} (t + 1)), $$  \hspace{1cm} (7)

where $N_r(i,j)$ is the $r$-neighborhood of cell $C(i,j)$ as $N_r(i,j) = \{ (k,l) \mid C(k,l) \text{ max } \{ |k-i|, |l-j| \} \leq r \}$. $x_{ij}(t)$, $y_{ij}(t)$ and $u_{ij}$ indicate the internal state, the output of cell, and the input of cell $C(i,j)$, respectively. The output function $f(\cdot)$ corresponds to the rounding operation. The output function plays an important role that the interpolation is optimized considering the nonlinearity caused by quantization noises. As shown in Figure 2, the image pixels are set to $u_{ij}$ and the equilibrium output $y'_{ij}$ is obtained after the transition of the network.

Next, the output of the first layer DT-CNN becomes the input of the second layer DT-CNN which has no dynamics. The output of the second layer DT-CNN provides the predicted values (see Figure 3). The image enhancement is composed of the horizontal and vertical processes illustrated in Figure 4. At the horizontal resolution enhancement stage, the pixel with coordinates $(i,j)$ of the input image is mapped to pixel $(i',j)$ of the resolution enhancement image which enlarged horizontally. Let $d_m$ be an enlargement parameter, the relationship between the pixel $(i,j)$ and $(i',j)$ is determined as $(i',j) = (id_m,j)$. The deficient pixel with coordinates $(k,l)$ of the enlarged image is obtained by

$$\hat{y}_{kl} = \sum_{y_{ij} \in N_r'(i',j)} B_h(i',j,k,l) y_{ij}',$$  \hspace{1cm} (8)

where $B_h$-template can be extended A-template of the first layer DT-CNN horizontally, that is,

$$B_h(i',j,k,l), \quad C(k,l) \in N_r'(i',j)$$

$$= \frac{1}{2 \sigma^2 \pi} \exp \left( -\frac{(k-i')^2}{2 \sigma^2} \right),$$  \hspace{1cm} (9)

$$N_r'(i',j) = \{ (k,l) \mid k-i' \leq d_m, \max \{ |l-j| \} \leq r \}.$$  \hspace{1cm} (10)

In the same manner, at the vertical resolution enhancement stage, the pixel $(i,j)$ is mapped to the pixel $(i,j') = (i,jd_m)$, and enlarged image is obtained by

$$\hat{y}_{ij} = \sum_{y_{ij} \in N_r'(i,j')} B_v(i,j',k,l) y_{ij}' ,$$  \hspace{1cm} (11)

where $B_v$-template obtained by

$$B_v(i,j',k,l), \quad C(k,l) \in N_r'(i,j')$$

$$= \frac{1}{2 \sigma^2 \pi} \exp \left( -\frac{(k-i')^2 + (l-j')^2}{2 \sigma^2} \right),$$  \hspace{1cm} (12)

$$N_r'(i,j') = \{ (k,l) \mid k-i' \leq r, \max \{ |l-j'| \} \leq d_m \}.$$  \hspace{1cm} (13)

Figure 2: First layer DT-CNN.
3. Experimental Results

We evaluated the effectiveness of our proposed image resolution enhancement method using the two-layered DT-CNN. We applied our method to the 8-bit gray-scale standard test images; “Aerial,” “Airfield,” “Boat,” “Couple,” “Crowd,” “Lena,” “Sailboat,” and “Tiffany” (Figure 5). Test images were resized from 512×512 to 256×256 by lowpass filtering and down sampling. Then, image resolution enhancement methods were applied.

The performance of the proposed method was compared with that of the bicubic interpolation algorithm which is indicated as “BC” and the wavelet interpolation algorithm using the well-known Le Gall 5/3 tap filter which is indicated as “WT”. For the simulation, the parameters were decided experimentally; the standard deviation of Gaussian $\sigma=0.54$, the $r$-neighborhood of cell $r=3$, the regularization parameter $\lambda=-1$. In order to compare with the resolution enhancement performance of WT, the enlargement parameter $d_m$ were decided as 2.

Table 1 shows the results of peak signal-to-noise ratio (PSNR) values between the original image and the interpolated image. The proposed method outperforms the BC and has better and competitive performances compared with WT.

Figure 6 shows a subjective comparison among the proposed method, BC, and WT for the “Lena” image with $d_m=2$. It is recognized that the jaggy-artifact is suppressed in comparison with WT.

Figure 5: Images used for the experimental evaluation of the proposed method (512×512 pixels): (a) Aerial, (b) Airfield, (c) Boat, (d) Couple, (e) Crowd, (f) Lena, (g) Sailboat, and (h) Tiffany.
Table 1: PSNR of resolution enhancement images (from 256×256 to 512×512)

<table>
<thead>
<tr>
<th>Method / Image</th>
<th>Aerial</th>
<th>Airfield</th>
<th>Boat</th>
<th>Couple</th>
<th>Crowd</th>
<th>Lena</th>
<th>Sailboat</th>
<th>Tiffany</th>
</tr>
</thead>
</table>

Figure 6: Detail of resolution enhancement images for Lena.

4. Conclusion

A novel image resolution enhancement method using two-layered DT-CNN was proposed. Our proposed method makes good use of the nonlinear interpolative effect of A-template to obtain an enhanced resolution image. And we can apply the arbitrary resolution enhancement parameter in contrast to the wavelet based method. The experimental results show that our proposed method outperforms the bicubic interpolation and has better and competitive performances compared with the conventional wavelet transform resolution enhancement methods.

Reference


Abstract—This paper illustrates a novel edge detection algorithm based on the Cellular Neural Network (CNN) paradigm. The approach exploits a rigorous model of the image contours and takes into account some electrical restrictions of existing hardware implementations. The analysis of benchmark video sequences (Car-phone, Stefan, and Foreman) highlights that the algorithm yields accurate results, better than the ones achievable by CNN-based methods and standard edge detection techniques.

1. Introduction

It is well-known that Cellular Neural Networks (CNNs) are nonlinear dynamical systems, constituted by locally interconnected cells, suitable for image processing applications [1]-[2]. The aim of this paper is to show that CNNs can be exploited for obtaining accurate edge detection. Note that edge detection is a crucial task in almost any image processing and computer vision application. Namely, edges convey essential information in a picture, and their fast detection is of primary importance during scene segmentation in video coding applications [3]. This paper focuses on CNNs since they enable real-time image processing to be obtained, by virtue of the great computation power offered by the CNN universal machine (CNNUM) [3]-[4]. More precisely, this paper presents a novel CNN-based edge detection algorithm, which exploits a two concentric circular windows operator for revealing edges as zero-crossing points of a difference function that depends only on the minimum and maximum values in the two windows. The approach, which consists of preliminary edge detection and final edge detection (see Section 3), is based on templates and parameters that take into account the constraints dictated by the hardware characteristics of the CNNUM. Namely, all templates and parameters have been designed by considering that the maximum allowed value for the template coefficients is 3, the upper bound for the bias values is 6, and the chip accuracy for the parameter values is 8-bit [5].

2. CNN model

Herein the standard CNN model is considered [6]:

\[
\dot{x}_i(t) = -x_i(t) + \sum_{u \in N_r} A_{u,i} y_u(t) + \sum_{u \in \omega_u} B_{u,i} u_u(t) + I_i(t)
\]  

(1)

where \(x_i(t)\) is the state, \(y_i(t)\) is the output, \(u_i(t)\) is the input, \(I_i\) is the bias, \(A_{u,i}\) and \(B_{u,i}\) are the parameters forming the feedback template \(A\) and the control template \(B\), respectively, whereas \(k l \in N_r\) is a grid point in the neighborhood within the radius \(r\) of the cell \(ij\). Five different images can describe a CNN layer, that is, the input \(U\), the state \(X\), the output \(Y\), the bias \(I\) and the mask \(M\). Finally, we recall that the CNNUM incorporates advanced computational capabilities, that is, \(\text{ADDITION/SUBTRACTION}\) on grayscale images, \(\text{AND/OR}\) on binary images and selection of the cells to be processed via the mask \(M\) [4].

3. CNN-based edge detection

3.1. Preliminary edge detection

From a theoretical point of view, an edge is a step or a slope between two uniform luminance areas. Unfortunately, edges in real scenes rarely comply with such statement. An efficient edge detector should comply with some basic requirements: (i) edges should be found with a low probability of false detection due to noise; (ii) edges should not be misplaced; (iii) algorithms should perform similarly with any image. The proposed approach exploits a basic concept illustrated by one of the authors in [7]. Let \(I(x,y)\) be a gray-level image, where \(I\) is the luminance signal. For each sample \(s\) belonging to \(I(x,y)\), let us consider two concentric circular windows, centered in \(s\) and having radius \(r\) and \(R\), respectively \((r<R)\). Let \(M^s\) and \(m^s\) be the maximum and minimum values of \(I\) within the window of radius \(R\), whereas let \(M'\) and \(m'\) be the maximum and minimum values of \(I\) within the window of radius \(r\) (see Fig.1). By taking \(\alpha_i(s) = M^s - M'\) and \(\alpha_i(s) = m' - m^s\), and by assuming that \(s\) is in the middle point of a luminance transition (in absence of noise), the relationship \(\alpha_i(s) = \alpha_i(s)\) holds [7]. In the case of noise, the change in the sign of the following difference function

\[
D(s) = \alpha_i(s) - \alpha_i(s)
\]

(3)

is a more suitable indicator of the presence of contours [7]. Note that \(D(s)=0\) is equivalent to find the flex points of
luminance transitions, since \( D(s) = \frac{\partial^2 I}{\partial n^2} \) approximates the directional derivative of the luminance \( I \) along the gradient direction [7].

Based on these considerations, we need to look for both the zero-points and the zero-crossing points of \( D(s) \). By introducing a threshold \( \mathcal{G} \), preliminary edge detection of samples \( s \) can be carried out as follows:

for each \( s = (x_s, y_s) \) so that \(-\mathcal{G} < D(s) < \mathcal{G}\) if \( D(s) = 0 \) \( \Rightarrow \) \( s \) is edge

elseif \( D(s) \geq 0 \) and

\[
\begin{align*}
D(x_s-1, y_s) < 0 & \text{ or } D(x_s+1, y_s) < 0 \\
\text{or } D(x_s, y_s-1) < 0 & \text{ or } D(x_s, y_s+1) < 0
\end{align*}
\]

\( \Rightarrow \) \( s \) is edge.

The operations described in (4) are now implemented using the CNN paradigm. We need to generate two images containing circular windows of radius \( r \) and \( R \), respectively, since the operations illustrated in Fig. 1 for a sample \( s \) have to be extended to the whole grey level image. Therefore, we generate two grid images, one containing only the circular windows of radius \( r \) (Fig. 2(a)) and the other containing only the circular windows of radius \( R \) (Fig. 2(b)).

We need to compute from the initial grey level image the four values \( M^8 \), \( m^8 \), \( M' \) and \( m' \). By starting from the top (left corner), these values have to be computed by means of hardware-oriented \textit{max} and \textit{min} templates (not available in literature). Referring to the \textit{max} templates, the proposed approach is based on two steps. The first one consists in the consecutive execution of eight \textit{rotated difference} templates:

\[
A_i = 0; \quad B_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}; \quad I_i = 0;
\]

\[
\begin{array}{cccc}
\ldots & \ldots & \ldots & \ldots \\
A_i = 0; & B_i = \begin{bmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}; & I_i = 0
\end{array}
\]

where the resulting image is the sum of the intermediate results. In the second step the output of the sequence (5) is supplied as bias map to the template \( A \) having the central element equal to 1 and the remaining elements zero (along with \( B = 0 \) and \( I = 0 \)). A similar process is used in order to realize the \textit{min} templates. By computing \( \alpha_r(s) \) and \( \alpha_s(s) \), a partial function \( D(s) \) is obtained. Namely, in order to find the final difference function \( D(s) \), it is necessary to iterate the previous operations. This can be done by moving the initial grey level image, along with those containing the two grids of circular windows, using the \textit{right} and \textit{down} templates reported in [6].

Once \( D(s) \) has been computed, we need to satisfy the condition \(-\mathcal{G} < D(s) < \mathcal{G}\). This can be done by applying the \textit{threshold} template reported in [6], along with the operations \textit{inversion}, \textit{OR} and \textit{inversion} again (via the CNNUM). Successively, by using the following \textit{increase} template:

\[ A = 0; \quad B = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}; \quad I = 4; \]

the conditions \( D(x_s \pm 1, y_s) < 0 \) or \( D(x_s, y_s \pm 1) < 0 \) are satisfied, whereas the \textit{AND} operation enables the constraint \( D(s) \geq 0 \) to be satisfied. Finally, by exploiting the OR operation, preliminary edge detection is obtained. Referring to \textit{Car-phone} and \textit{Stefan}, which are two of the benchmark sequences in video coding applications, the results of the preliminary phase are shown in Fig. 3.
3.2. Final edge detection

The zeros of $D(s)$ are not only flex points of luminance transitions, but also the set of pixels where luminance is almost constant. Owing to noise, small fluctuations may result in changes in the sign of $D$ that would be incorrectly assumed as edge points (see both Fig.3(b) and Fig.3(d)). Therefore, in order to better select the previously detected edges, we propose to derive from $D(s)$ an image (called $P(s)$) able to highlight the discontinuity areas. By using $P(s)$, we can first obtain an image including all the contours selected by the gradient operation. Then, final edge detection is achieved by making the open contours closed.

Previous operations are now implemented using the CNN paradigm. At first the image representing $D(s)$ is processed by means of the threshold template reported in [6], with $I_0=0$. Then, the pixels in $D(s)$ that correspond to $D(s)>0$ assume the values $R_M$, whereas the pixels that correspond to $D(s)<0$ assume the values $R_m$. The resulting two images are indicated with $D(s)R_M$ and $D(s)R_m$, respectively. By means of the zero* template reported in [4], the two images $D(s)R_M$ and $D(s)R_m$ are combined with the result of the threshold operation. The resulting images $P(s)$ for Car-phone and Stefan video sequences are reported in Fig.4.

![Fig. 4 Images representing P(s): Car-phone (a); Stefan (b).](image)

Then, by combining the binary mask of $P(s)$ with the output of the preliminary phase via the zero* template, we obtain the images reported in Fig.5, which include all the contours selected by the gradient operation.

![Fig.5 Contours selected by the gradient: Car-phone (a); Stefan (b).](image)

In order to reduce all the contours shown in Fig.5 to one-pixel thin lines, the skeletonization process is applied. The proposed hardware-oriented approach exploits eight linear templates, in the eight directions N, NW, NE, W, E, SW, SE and S. For example, in the NW direction the proposed novel templates are:

\[
A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0.25 & 0.25 & 0 \\ 0.25 & -0.25 & -0.25 \\ 0 & -0.25 & 0 \end{bmatrix} \quad I = -0.75. \quad (7)
\]

Then, in order to make the open contours closed, the following novel sequence of closing templates is suggested in the NW direction:

\[
A_1 = 0, \quad B_1 = \begin{bmatrix} -0.1 & -0.1 & -0.1 \\ -0.1 & 3 & 0.1 \\ 0.1 & -0.1 & -0.1 \end{bmatrix}, \quad I_1 = 1
\]

\[
A_2 = 0, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0.1 \\ 0.1 & -0.1 & -0.1 \end{bmatrix}, \quad I_2 = 0. \quad (8)
\]

The images showing the final contours for Car-phone and Stefan video sequences are reported in Fig.6, whereas the final edge detection for Foreman video sequence is shown in Fig.7.

![Fig.6 Final edge detection: Car-phone (a); Stefan (b).](image)

![Fig.7 Foreman: sample frame (left); final edge detection (right).](image)

4. Discussion

Now the CNN-based edge detection method illustrated in [3] and the proposed one are analyzed and discussed. Referring to Car-phone, the results are reported in Fig.8.

![Fig. 8 Car-phone video sequence: (a) edge detection by the method in [3]; (b) edge detection by the proposed approach.](image)
The comparison clearly highlights the effectiveness of the proposed technique. It can be observed that our approach presents the following features: i) contours are always single-pixel and are much more close to the real contours; ii) only closed contours are obtained (by virtue of the closing templates (8)). The first feature is important because in object-oriented coding schemes the moving regions should coincide with the moving contours of a real image [3]. Namely, in the context of object-oriented coding schemes, the main task to be accomplished is the segmentation of a scene into different moving regions. These regions should coincide with the moving contours of a real image in order to guarantee that the description of the motion is efficient, since errors resulting from bad edge detection are highly visible to the human eye and significantly decrease the quality of the video sequence. Referring to the second feature, closed contours are better over open ones because only closed contours enable moving objects to be detected and extracted from a scene in a video sequence. This is a key point of our method, since it enables closed regions with uniform luminance to be effectively detected, which is of paramount importance when dealing with object-oriented coding and object recognition [7].

4.1. Comparisons with standard edge detection techniques

Comparisons are carried out between standard edge detection methods (i.e., methods not involving CNNs) and the proposed CNN-based approach. The techniques considered herein are the well-established LoG edge detector [8] and Canny algorithm [9]. Referring to Car-phone, the result using the LoG (Laplacian of the Gaussian) method is reported in Fig.9(a), whereas the outcome of the Canny edge detector is shown in Fig.9(b), both with $\sigma = 1.2$ Gaussian filtering.

Fig. 9 Edge detection using standard techniques: (a) result using the LoG; (b) result using the Canny algorithm.

The comparison between the CNN-based approach reported in Fig.8(b) and the standard techniques shown in Fig.9 highlights, once more, the key point, that is, only closed contours are generated by our method. This is of great importance in view of the development of segmentation algorithm for object-extractions.

4.2. Evaluation of the processing time

An estimation of the processing time achievable by the proposed method is given. As in [4], we are able to estimate an execution time of $10716 \tau$ for Car-phone video sequence. By taking $\tau = 280$ns [4], we obtain a processing rate of about 333 frame/s, which greatly exceeds the usual video frame of 30 frame/s. Similar processing rates, for other video sequences, confirm the great capability offered by the proposed technique.

5. Conclusions

Accurate and fast edge detection is a crucial task in almost any image processing and computer vision application. In this paper we have proposed an effective CNN-based edge detection technique, which can be useful in the context of real-time segmentation for object-oriented video coding applications. The approach, based on rigorous model of the image contours, has exploited some novel templates that can run on the CNNUM. The analysis of benchmark video sequences has shown that the method yields very accurate results, better than the ones achievable by standard techniques and CNN-based methods.

References

Measuring the skin surface changes due to hydrating treatments through capacitive images analysis

Alessandro Bevilacqua and Alessandro Gherardi

ARCES (Center of Excellence on Electronic Systems)
University of Bologna, Via Toffano 2/2, 40125 Bologna, ITALY
Email: {abevilacqua, agherardi}@arces.unibo.it

Abstract—Quantitative evaluation of the changes in skin topographic structures are of a great importance in the dermocosmetic field to assess subjects response to medical or cosmetic treatments. Although many devices and methods are known to measure these changes, they are not suitable for a routine approach and most of them are invasive. We describe a non-invasive method based on a portable capacitive device to assess short term changes of the topographic structures of the skin surface. This is based on a feature we have developed and already used to measure long-term changes due to the ageing process. We measure the difference between the feature values computed before and after a hydrating treatment in different aged subjects. Experiments show a reduction of the overall amount of micro-wrinkles which is more than twice in young subjects with respect to the older ones.

2. Previous Work

Follow up studies generally use photographic images of the treated skin surface to assess differences according to a scoring or a scaling system. Despite the variety of published scoring and scaling systems for assessing skin topographic changes, none of them has been established as being the standard. Moreover, most of the methods either rely on visual evaluation or achieve quantitative evaluation by means of histologic examinations, which are invasive techniques.

Many devices are known to measure the skin topographic age-related changes, but only few of them are portable and suitable for a routine approach. Works in [3, 1] describe a method to extract topographic features linearly correlated with the skin ageing process by using a portable capacitive device. Two topographic structures of the skin are analysed. In [3] a feature related to the skin cells (polygonal patterns identified by micro-wrinkles) area distribution has been found to be correlated to skin ageing, while in [1] a feature related to wrinkles, the mean of the local contrast of the wrinkle enhanced image, has been successfully employed to measure skin ageing.

As stated before, image mosaicing is necessary to obtain a large image starting from single acquisition images. Image mosaicing and therefore image registration, presents large number of applications in digital imaging. In the past decades many solutions have been presented for merging images from aerial views, video sequences and documents. Few methods have been devised for medical images where maintaining the original resolution is essential as well as allowing translation and rotation transforms. For example, the very interesting method described in [5] deals with histological skin images. It estimates the amount of overlap between images by ensuring the sub-pixel accuracy with phase correlation. However, although it can deal with
large displacements, it covers only translational displacements, by forcing the acquisition to exclude any rotation.

3. The Device used

To achieve a digital representation of the skin surface, a capacitive device has been used. This device, originally developed for fingerprint acquisition and recognition in biometric applications, relies on an active capacitive pixel sensing technology and it is able to capture detailed images of the skin at a resolution of 50 µm/pixel. The sensor’s array is composed of $256 \times 360$ pixels and is of $12.8 \times 18.0$ mm. The surface of each pixel is composed of two adjacent metal plates (Fig. 1). When live skin is brought near to the sensor plates, the skin interferes with field lines between the two plates and the feedback capacitance is minimized. Conversely, when the skin is moved away from the sensor surface the feedback capacitance is maximized. Therefore, the gray level values of the output image represent the 3D skin surface: darker under skin tissue and lighter under skin wrinkles. Images are acquired in real-time by pressing the device lightly over a skin’s Region of Interest (ROI).

4. What we measure and how

This work presents a method to assess short term changes of the topological structures of the skin surface of different subjects treated with a hydrating cream. Firstly, we extract in two subsequent acquisitions (before and after treatment) the same skin area. Among the skin macro structures, we have focused our attention on skin micro-relief, which wide as age increases. Skin micro-relief is composed of many micro-wrinkles and few larger wrinkles. The effect of the hydrating treatment is mostly visible for micro-wrinkles, which are relaxed and sometimes disappear. Secondly, for each acquisition we give a measure related to the overall amount of wrinkles by using the feature described in [1]. The difference between the two measures (given in percent) represents the output of our system.

4.1. The feature analysed

To achieve a better representation of the skin topographic structure of interest, different processing steps are needed. In particular, noise and different pressure levels applied to the sensor surface affects the gray level value of the capacitive images acquired. In order to make the images coming from different subjects and body sites comparable to each other, image normalization is required. To this purpose, a local contrast enhancement is applied before any subsequent feature extraction. The MLC feature we have devised is based on the evaluation of the amount of wrinkles detected by the capacitive device. From a topographic point of view, wrinkles can be approximately considered as continuous curves joining line segments having different spatial orientations. Therefore, a line enhancement algorithm is used to emphasize wrinkles. To this purpose, we use a set of bidimensional Gaussian filters having 14 different orientations. After that the image has been filtered using the Gaussian filtering, we have a set $E$ of filtered images, one for each orientation which the filter has been applied for. Each of these images retains all the wrinkles components (in pixels) along one specified direction. Therefore, what we measure by MLC is the wrinkles component of each capacitive image, represented by the mean value of the local contrast expressed by Eq. 1:

$$MLC = \frac{\sum_{(x,y) \in E} \sigma_w(x, y)}{\text{card}(E)}$$  \hspace{1cm} (1)

where $\sigma_w(x, y)$ is the standard deviation value of the pixel $(x, y)$ over a window of size $w$ centered on the point $(x, y)$ for each Gaussian filtered image in the set $E$. Practically speaking, MLC value is proportional to the overall amount of area occupied by wrinkles. This feature, validated on 90 subjects, has shown a linear correlation with the subjects age, thus proving the efficacy of the method to measure wrinkles topographic changes.[1] Since MLC increases with age, the differential analysis should indicate a reduction in its value due to the effect of the hydrating treatment.

In this work, the differential analysis relies on the MLC, which is measured before and after the cream application, thus allowing to quantify the effect of the hydrating treatment.

4.2. Image acquisition

In order to perform a reliable differential analysis we need to perform two subsequent acquisition, at different time (hours or days as well), sharing (part of) the same body site. Since for obvious reasons employing artificial markers is not feasible, we could only rely on human guidance and natural markers. However, this approach would produce a shared area smaller than the sensor surface and it would be not enough to perform our measures reliably. We overcome this problem by acquiring more images for each acquisition session covering a large area and building a global mosaic. Accordingly, we define the acquisition session as the whole set of acquisition shots performed on a same skin region in a reduced amount of time (less than
one minute). This permits to consider skin physiological properties as being constants and to cover, at sight, an area large enough to perform a reliable differential measure.

4.3. Automatic Image Mosaic

The image registration is the process of aligning two images of the same scene by warping the first one (input image) into the second one (base image).

To estimate the transformation between overlapping images we need firstly to extract a set of interesting points from both images to search for correspondences between them. As the feature point extractors we use Harris [6] because it has proved to be the better among those considered. Since the skin is an elastic and deformable medium, the pressure applied manually by contact during acquisition yields elastic deformation of the skin. This necessarily influences the complexity of the transformation model to choose. By relaxing the constraint of elastic deformation, we use an affine transformation model, which is a 6-parameter linear combination of translation, rotation and scaling. It permits different stretching along image rows and columns and it is the most general linear transformation. For a robust assessment of the affinity between the two images, we use the RANSAC [7] algorithm which simultaneously computes the affinity and rejects the outliers. Fig. 2 shows two images of the upper ventral forearm region of a 25 y.o. male successfully registered and blended in the resulting mosaic image. As we can see the stitching is seamless and the skin structure of both images is preserved. This is the simplest registration case involving two images. As for the registration of more than two images, we perform an all-to-all registration which produces a graph with skin images at the nodes and the corresponding affinities. Computing the minimum spanning tree of the graph, leads to reduce the registration errors that may arise from concatenating affinities. This permit to obtain a common area between two different acquisition sessions large enough to achieve reliable measures for differential analysis.

5. Experimental results

This work aims at quantify short term changes of the topographic skin structures of 30 subjects belonging to two classes: 20-30 y.o. and 50-60 y.o., respectively. The experiments are accomplished on the upper ventral forearm region, which is the least suffering from environmental exposure. The results attained strongly depend on life habits (e.g. usually, women use more beauty care products than men) and on the physiological response of each subject to hydrating treatments. Choosing far classes allows to appreciate in a better way how ageing affects the skin elasticity. In this work, we analyse a young man and a young woman belonging to the first class and an older woman belonging to the second class. The first acquisition session is performed before applying any treatment. Further on, a hydrating cream is applied and after one hour the effects of the hydrating treatment become visible. Then we perform the subsequent acquisition sessions and extract the shared area from the two mosaics. Fig. 3 shows two images extracted from the shared area of the two mosaics referring to a 22 y.o. female, before (left) and after (right) treatment.

The MLC values have been calculated before and after the hydrating treatment. For all samples the value of the MLC feature after treatment decreases, thus proving an overall improvement. Fig. 4 shows the improvement for each subjects due to the hydrating treatment. As for the 22 y.o. female of Fig. 3, the improvement has been measured in 4.3%. You can see that for younger subjects the reduction of the skin micro-wrinkles is more than twice with respect to the older subject. This is due to the skin of the young subjects being more elastic and reactive than the skin of older subjects. In Fig. 5 you can see a common area ex-
Figure 4: MLC improvement after the hydrating treatment for the three subjects.

Figure 5: Shared area from two mosaics referring to a 54 y.o. female, before (left) and after (right) treatment.

tracted from two mosaics referring to the 54 y.o. subject, before (left) and after (right) hydrating treatment. The improvement is of 2.0%. Comparing this couple of images with the ones of Fig. 3 it becomes evident how the younger subject shows a better response to the hydrating cream application. In regard to the improvement for the 25 y.o. male, it is of 5.0%, which is higher than the one of the 22 y.o. female. Probably, this is due to the overall skin hydration state of the female subject being better than the male one. In fact, the 22 y.o. woman often applies herself hydrating creams, thus reducing the margin for further improvements. At the opposite, this is the first time for the 25 y.o. male to apply himself a hydrating cream, and his skin owns a higher absorption potential.

6. Conclusion and Future Work

This work describes a method to measure changes of the skin topographic structures due to hydrating treatments. Our approach relies on a portable capacitive device and on a feature which has been previously used to measure age-related changes of the skin surface. The experiments we accomplished show that this feature is effective also for changes due to different hydration. It is worth remarking that this work represents the first attempt of measuring such a kind of changes by using a portable capacitive device.

At the moment, we have started a follow up study involving 30 subjects for evaluating middle/long term changes due to hydrating treatments.

References


Learning Algorithm for Multiple Vector Quantization Based Image Compression

Noritaka Shigei †, Hiromi Miyajima † and Michiharu Maeda ‡

†Department of Electrical and Electronics Engineering, Kagoshima University
1-21-40 Korimoto, Kagoshima 890-0065, Japan
‡Kurume National College of Technology, Kurume Fukuoka 830-8555, Japan
Email: shigei@eee.kagoshima-u.ac.jp

Abstract—In image compression, multiple vector quantization with index inference (MVQII) can create a high quality image by combining multiple temporal restored images. However, MVQII trains codebooks so as to optimize not the final restored image but temporal restored ones. This paper proposes a learning algorithm for MVQII, which is referred as MVQII with Learning (MVQIIL). Unlike the conventional one, the objective function of MVQIIL is the error between the original image and the final restored one. The effectiveness of MVQIIL is shown by numerical simulations.

1. Introduction

Adaptive Vector Quantization (AVQ) is to find a small set of weight vectors that will approximate a larger set of input vectors[1, 4]. Various neural network algorithms for AVQ have been proposed[3, 4]. AVQ is a useful technique in lossy data compression such as image compression[2, 4, 5, 6, 7, 8]. When using AVQ in image compression, the restored image quality greatly depends on the number of weights (codebook size) and the dimension of weights (block size). One of the methods relaxing the trade-off is Multistage Residual VQ (MRVQ), in which multiple quantizers are concatenated in series [2]. In MRVQ, the first stage quantizer operates on the input vectors, and the second stage one operates on the errors between the input vector and the first stage output. MRVQ is a serial approach. On the other hand, we have proposed multiple-VQ (MVQ) as a parallel approach[9]. The MVQ method performs independently multiple VQ processes, and combines the independent low quality results into a high quality one. We have shown that, with a useful technique index inference, MVQ is effective in image compression. However, MVQ with index inference (MVQII) trains the codebooks so as to optimize not the final restored image but temporal ones. This fact implies that there is much to be improved in this approach.

In this paper, we propose a learning algorithm for MVQII, which is referred as MVQII with Learning (MVQIIL). Unlike the conventional one, the objective function of MVQIIL is the error between the original image and the final restored one. Numerical simulations are performed to investigate effective parameter combinations for MVQIIL and show the effectiveness of MVQIIL. The simulation results show that 1) MVQIIIL achieves 6 \sim 10\% improvement from MVQII, and 2) when the compression rate is high, MVQIIIL outperforms single VQ, MRVQ and MVQII.

2. AVQ and VQ Based Image Compression

Vector Quantization (VQ) is to approximate a large set of input vectors \( X = \{x_1, \ldots, x_n\} \) by a smaller set of weight vectors \( W = \{w_1, \ldots, w_n\} \), where \( x_i, w_i \in \mathbb{R}^n \) are \( n \)-dimensional Euclidean vectors and \( X \) is a random sample from a probability density function (PDF) \( p(x) \). In VQ, an input vector \( x \in X \) is replaced with a weight vector \( w_i \in W \) such that \( d(x, w_j) = \min_{w \in W} d(x, w) \), where \( d(x, w) \) is the squared error \( |x - w|^2 \). In other words, the VQ procedure divides the input space \( \mathbb{R}^n \) into \( \kappa \) subspaces \( S_1, \ldots, S_\kappa \) such that \( S_i = \{x \in \mathbb{R}^n | d(x, w_j) \leq d(x, w_j), j \neq i\} \). The approximation accuracy of VQ is evaluated in terms of the average distortion error \( E = \frac{1}{n} \sum_{x \in \{1, \ldots, \kappa\}} \sum_{S \in S} d(x, w_j) \). If \( p(x) \) is well-known, optimal weight vectors minimizing \( E \) can be created directly from \( p(x) \). When \( p(x) \) is not well-known, an Adaptive Vector Quantization (AVQ) procedure estimates \( W \) from \( X \). The estimation can be performed by using neural network algorithms such as competitive learning, neural-gas and Kohonen’s self-organizing map[3, 4].

A fundamental method minimizing the error \( E \) is Competitive Learning (CL), which is based on gradient descent[4]. The CL procedure iterates a simple adaptation step. At \( t \)-th iteration, the CL procedure calculates the closest weight (called \textit{winner}) \( w_{\text{win}} \) to a given input vector \( x \in X \), and then updates \( w_{\text{win}} \) as follows:

\[
\begin{align*}
  w_{\text{win}} & \leftarrow w_{\text{win}} + \varepsilon(t)(x - w_{\text{win}}),
\end{align*}
\]

where \( d(x, w_{\text{win}}) = \min_{j \in \{1, \ldots, \kappa\}} d(x, w_j) \), and \( \varepsilon(t) \) is a learning rate decreasing with \( t \).

2.1. VQ Based Image Compression

This section describes single-VQ based image compression (SVQ). We assume \( G \)-bit gray-scale images of \( M \times N \)
pixels. Let \( p_{i,j} \in \{0, \cdots , 2^G - 1\} \) \((i \in \{1, \cdots , M\}, j \in \{1, \cdots , N\})\) be the gray level of the pixel at coordinates \((i,j)\). Then an image is represented as an \( M \times N \) matrix \( P = (p_{i,j}) \). A typical algorithm of SVQ is as follows.

**Algorithm SVQ**

**Step 1 (Input Preparation)**

Given an input image \( P \). The \( M \times N \) pixels in \( P \) are divided into \((M \times N)/(J \times K)\) blocks of size \( J \times K \). The blocks are represented as \( J\)-\( K\)-dimensional vectors \( x_1, \cdots , x_{(M \times N)/(J \times K)} \).

**Step 2 (Vector Quantization)**

The set of weight vectors \( W = \{w_1, \cdots , w_N\} \), called codebook, is trained by using the set of vectors \( X = \{x_1, \cdots , x_{(M \times N)/(J \times K)}\} \) as input data. The training is performed by a neural network AVQ algorithm such as Eq.(1).

**Step 3 (Index Calculation)**

For each \( i \in \{1, \cdots , (M \cdot N)/(J \cdot K)\} \), the index number \( l_i \) is calculated, where \(|x_i - w_{l_i}| = \min_{j \in \{1, \cdots , N\}} |x_i - w_j|\).

**End of SVQ.**

The compressed data is composed of the codebook \( W \) and the index sequence \( L = (l_1, \cdots , l_{(M \cdot N)/(J \cdot K)}) \). From \( W \) and \( L \), an image is restored. In the restored image, each block \( x_i \in X \) in the original image is replaced with \( w_{l_i} \).

### 3. Multiple-VQ Based Image Compression

#### 3.1. MVQ and Index Inference

In this section, we describe multiple-VQ based image compression (MVQ). The fundamental idea of MVQ is that a high quality image can be created by averaging multiple low quality images independently restored from different pairs of codebook and index sequence. The MVQ algorithm for compression phase is as follows.

**Algorithm MVQ**

**Step 1 (Input Preparation)**

From the input image \( P \), two input data sets \( X^{(0)} \) and \( X^{(1)} \) are generated as follows:

\[
X^{(c)} = \{x^{(c)}_i|i \in \{1, \cdots , MN/j_cK_c\}\} \quad \text{for} \quad c \in \{0,1\}, \\
x^{(c)}_i = (p_{\phi(i,c,1),\psi(i,c,1)}; \cdots ; p_{\phi(i,c,J_c),\psi(i,c,1)}; \cdots ; p_{\phi(i,c,J_c),\psi(i,c,2)}; \cdots ; p_{\phi(i,c,J_c),\psi(i,c,K_c)}), \quad i \in \{1, \cdots , MN/(J_cK_c)\},
\]

where \( \phi(i,c,j) = ((i-1) \mod \frac{M}{J_c})J_c + c + j \mod M \) and \( \psi(i,c,j) = ((i-1)/J_c)K_c + c + j \mod N \). Fig. 1 is a schematic explanation of \( x^{(c)}_i \).

**Step 2 (Vector Quantization)**

Perform VQ twice to generate two codebooks \( W^{(0)} \) and \( W^{(1)} \), where codebooks \( W^{(0)} \) and \( W^{(1)} \) are trained by using data sets \( X^{(0)} \) and \( X^{(1)} \), respectively. The training is performed by a neural network AVQ algorithm such as Eq.(1).

**Step 3 (Index Calculation)**

The index sequence \( L^{(0)} = (l^{(0)}_1, \cdots , l^{(0)}_{(MN)/(J_0K_0)}) \) is calculated from \( W^{(0)} \).

**End of MVQ.**

In Step 2, two codebooks are generated from different input data sets. The use of different data sets provides a better performance than the one of the same data set[9].

In Step 3 only \( L^{(0)} \) is calculated, and the compressed data are \( W^{(0)} \) and \( W^{(1)} \). If, in addition to the compressed data, \( L^{(1)} \) is used, a better quality image can be restored. However, in this case, the compressed data size significantly increases, and as a result, the restored image quality is too low for the data size. Instead, the lost data \( L^{(1)} \) is restored by a decoder side technique index inference[10]. The algorithm is as follows:

**Algorithm Index Inference**

**Step 1 (Image Restoration by SVQ)**

Restore an image \( \tilde{B}^{(0)} = (\tilde{p}_{i,j}^{(0)}) \) from \( W^{(0)} \) and \( L^{(0)} \).

**Step 2 (Input Preparation for Inference)**

From the restored image \( \tilde{B}^{(0)} \), input data set \( \tilde{X}^{(1)} = \{\tilde{x}^{(1)}_1, \cdots , \tilde{x}^{(1)}_{(MN)/(J_1K_1)}\} \) is generated as in Eq.(2).

**Step 3 (Index Calculation)**

For each \( i \in \{1, \cdots , (MN)/(J_1K_1)\} \), the index number \( l^{(1)}_i \) is calculated, where \(|\tilde{x}^{(0)}_i - w_{l^{(1)}_i}^{(1)}|^2 = \min_{j \in \{1, \cdots , N\}} |\tilde{x}^{(0)}_i - w^{(1)}_j|^2|\).

**End of Index Inference.**

Now, after index inference, we have \( W^{(0)} \) and \( L^{(0)} \) and \( L^{(1)} = (l^{(1)}_1, \cdots , l_{(MN)/(J_1K_1)}) \). Next, two images
\( \tilde{P}^{(0)} \) and \( \tilde{P}^{(1)} \) are independently restored from two pairs \( (W^{(0)}, L^{(0)}) \) and \( (W^{(1)}, L^{(1)}) \), respectively. And then, a restored image \( \tilde{P} = (\tilde{p}_{i,j}) \) is generated by combining the two restored images as follows:

\[
\tilde{p}_{i,j} = \frac{1}{2} \left( \tilde{p}_{i,j}^{(0)} + \tilde{p}_{i,j}^{(1)} \right),
\]

where \( \tilde{P}^{(c)} = (\tilde{p}_{i,j}^{(c)}) \) for \( c \in \{0, 1\} \).

In the following, the method described in this section is referred as MVQ with Index Inference (MVQII).

### 3.2. Learning for MVQII

The codebooks should be trained so as to minimize the error between the original image and the final restored one. However, MVQII trains a codebook so as to minimize the error between the original image and a temporal restored image. The proposed algorithm MVQII with Learning (MVQIIL) is as follows:

**Algorithm MVQII with Learning**

**Steps 1 and 2**
These steps are same as Steps 1 and 2 in Algorithm MVQ.

**Step 3**
Set \( t \leftarrow 1 \), where \( t \) is the number of learning iterations.

**Step 4 (Index Calculation)**
Calculate \( L^{(0)} = \{ l^{(0)}_1, \ldots, l^{(0)}_{MN}/(J_0K_0) \} \) from \( W^{(0)} \) and \( X^{(0)} \).

**Step 5 ~ 7 (Index Inference)**
These steps are same as Steps 1 ~ 3 in Algorithm Index Inference.

**Step 8 (Input Preparation)**
Let \( c = t \mod 2 \). If \( c = 2 \), then go to Step 9. Otherwise, restore an image \( \tilde{P}^{(1)} = (\tilde{p}_{i,j}^{(1)}) \) from \( W^{(1)} \) and \( L^{(1)} \), and generate input data set \( \tilde{X}^{(0)} \) from \( \tilde{P}^{(1)} \).

**Step 9 (Calculating Correction Values)**
For each \( i \in \{1, \ldots, k_c\} \), \( \Delta w_{i} \leftarrow 0 \) and \( s_i \leftarrow 0 \). And then, for each \( i \in \{1, \ldots, (MN)/(J_0K_0)\} \),
\[
\begin{align*}
\Delta w^{(c)}_{i} & \leftarrow \Delta w^{(c)}_{i} + x^{(c)}_{i} - \frac{1}{2} (\tilde{x}^{(c)}_{i} + w^{(c)}_{i}), \\
s^{(c)}_{i} & \leftarrow s^{(c)}_{i} + 1.
\end{align*}
\]

**Step 10 (Updating Weights)**
For \( i \in \{1, \ldots, k_c\} \),
\[
w^{(c)}_{i} \leftarrow w^{(c)}_{i} + \frac{\Delta w^{(c)}_{i}}{s_i}.
\]

If \( t = T_{\text{max}} \), then the algorithm terminates. Otherwise, \( t \leftarrow t + 1 \) and go to Step 5.

**End of MVQ with Learning**

In Step 10, codebooks \( W^{(0)} \) and \( W^{(1)} \) are alternatively updated so as to minimize the error between the original image \( P \) and the restored one \( \tilde{P} \). Specifically, when \( t \) is odd (even), \( W^{(1)} \) \( (W^{(0)}) \) is updated. Empirically, we know that updating simultaneously \( W^{(0)} \) and \( W^{(1)} \) does not work better than this alternative fashion.

### 4. Numerical simulations

We perform numerical simulations to investigate effective parameter combinations for MVQIIL and show the effectiveness of MVQIIL. Fig.2 shows two test images used in the simulations. The simulations evaluate four methods: SVQ, MRVQ, MVQII and MVQIIL in terms of Means Square Error (MSE) between an original image and a restored image. For all the methods, AVQ uses a sufficient large number of iterations so as to minimize MSE. MVQII and MVQIIL are performed with \( J_0 \times K_0 = 4 \times 4 \), because the compression rate can be kept high (specifically up to 20.0). MVQIIL is performed with \( T_{\text{max}} = 6 \), which is sufficient large to converge MSE.

First, we search for effective combinations of block size \( J_1 \times K_1 \) and codebook sizes \( \kappa_0 \) and \( \kappa_1 \). We perform MVQIIL with different combinations of \( J_1 \times K_1 \), \( \kappa_0 \) and \( \kappa_1 \). For \( J_1 \times K_1 \), five cases \( 2 \times 1, 2 \times 2, 4 \times 1, 4 \times 2 \) and \( 4 \times 4 \) are tested. For \( \kappa_0 \) and \( \kappa_1 \), four cases 64, 128, 256 and 512 are tested. Fig.3 shows the simulation results for Lenna. The results clearly indicate that the large block size \( J_1 \times K_1 = 4 \times 4 \) degrades the performance. One of its reasons is that the data size of codebook \( W^{(1)} \) significantly increases with \( J_1 \times K_1 \). The effective combinations found in the simulations are summarized in Table 1.

Next, the proposed method MVQIIL is compared with SVQ, MRVQ and MVQII. MRVQ is performed by a software package QcPack[6]. For MRVQ, the number of stages is two and each stage uses the same codebook size. MVQIIL and MVQII are performed with the effective parameters shown in Table 1. Figs.4 and 5 show the simulation results for Lenna and Goldhill, respectively. Compared with MVQII, MVQIIL achieves approximately 10% and 6% improvements for Lenna and Goldhill, respectively. For both images, MVQIIL achieves the best performance among all the methods when the compression rate is high.

In data compression, VQ is often used with orthogonal transformation such as DCT, DWT and DFT[5, 7]. When solely using VQ, it would be hard to achieve a better perfor-

---

**Figure 2: Test images \((M = N = 512 \text{ and } G = 8)\).**

(a) Lenna  
(b) Goldhill
Table 1: Effective combinations of $J_1 \times K_1$, $\kappa_0$, and $\kappa_1$ and the compression rate for each combination.

<table>
<thead>
<tr>
<th>$J_1 \times K_1$</th>
<th>$2 \times 2$</th>
<th>$2 \times 2$</th>
<th>$4 \times 1$</th>
<th>$4 \times 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\kappa_0$</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>512</td>
</tr>
<tr>
<td>$\kappa_1$</td>
<td>64</td>
<td>128</td>
<td>256</td>
<td>256</td>
</tr>
<tr>
<td>Compression rate</td>
<td>19.3</td>
<td>15.5</td>
<td>12.2</td>
<td>9.5</td>
</tr>
</tbody>
</table>

Performance than existing effective compression methods such as JPEG. However, the development is still an important effort to realize high performance data compression.

5. Conclusions

In this paper, we propose a learning algorithm MVQIIL, which unlike the conventional one, its objective function is the error between the original image and the final restored one. According to the performed simulation results, 1) MVQIIIL achieves $6 \sim 10\%$ improvement from MVQII, and 2) when the compression rate is high, MVQIIL outperforms single VQ, MRVQ and MVQII.

References

An Efficiently Adaptive Ray Tracing by Automatic Differentiation

Masato Misumi† and Mitsunori Makino†

†Department of Information and System Engineering, Chuo University
1-13-27 Kasuga, Bunkyo-ku, Tokyo 112-8551, Japan
Email: mmisumi@makino.ise.chuo-u.ac.jp, makino@m.ieice.org

Abstract—The ray tracing algorithm is well-known as one of the most popular rendering techniques in computer graphics. In the ray tracing, tracing the rays and calculating intersection between rays and objects are essential. Since usually a ray having no thickness is defined for each pixel of the image, the image may include aliasing. For anti-aliasing, one of the authors proposed an adaptive distributed ray tracing in the use of the automatic differentiation. However, number of initial rays are automatically determined by resolution of prospective image. Therefore it takes longer time as we need larger image. Expanding the method to detection of variation over pixels, an adaptive method is proposed to determine from which pixels initial rays should be radiated. Since unnecessary initial rays are not calculated, the proposed method benefits more efficient calculation of ray tracing.

1. Introduction

Ray tracing[1] is well known as one of the most popular rendering techniques in computer graphics technology, which can generate photorealistic images by tracing rays from the given viewpoint. In the ray tracing, tracing the rays and calculating intersection between rays and objects are essential for us to generate images.

Since a thickness of ray is initially defined for each quadratic pixel on an image, resolution of the image is important to keep good quality of the image. Suppose that resolution, i.e., number of pixels constituting an image, is low in comparison with a given scheme. Then the image may include aliasing (jaggy) on some pixels, although it takes shortly for computation of the image. Otherwise if resolution is set to high, then generation of the image requires long computation time. It is very hard to determine appropriate resolution uniformly, since such resolution is influenced by objects, lights and viewpoints. Therefore an adaptive ray tracing technique is required, in which high resolution is applied to area corresponding to complex situation and low resolution is set for simple situation.

Cook proposed a distributed ray tracing[2], in which some rays are defined for each pixel to get information on neighbor area. While the method contributes anti-aliasing effect, it requires several fold computation time. Furthermore some pixels may need single ray for each when they correspond to very simple situation (e.g., no object).

For improving efficiency, Makino proposed an adaptive distributed ray tracing in the use of the automatic differentiation[3], which can detect edges of shapes and shadows and smooth them. By the automatic differentiation, neighbor situation around a pixel can be obtained as primary approximation. If the approximation is different from the result of the original ray, there may exist aliasing around the pixel. Then, distributing rays at the pixel, anti-aliasing effect is expected. The method makes the conventional distributed ray tracing adaptive. Since pixels having aliasing are relatively small among all pixels on an image, the method is usually more efficient than the conventional distributed ray tracing. Also application of the method to animation was shown, in which motion blur is efficiently detected[4].

However, number of initial rays are automatically determined by resolution of prospective image. Therefore it takes useless computational cost as we need larger image. Expanding the method to detection of variation over pixels, in this paper, an adaptive method is proposed to determine from which pixels initial rays should be radiated. Since unnecessary initial rays are not calculated, the proposed method benefits more efficient calculation of ray tracing.

2. An Efficiently Adaptive Distribution of Initial Rays

2.1. Problems to be Solved

The previously proposed method is applied to detection of pixels to be oversampled. In this paper we expand this technique to determination whether an initial ray from given viewpoint through a pixel should be traced or it can be approximated by neighbor rays. In this case, the followings must be considered:

1. Since distance between pixels is much longer than distance between center and edge of a pixel, the approximation by the automatic differentiation becomes less accurately.

2. In the distributed ray tracing, color (brightness) of a pixel is determined as arithmetic mean of color of its subpixels. Such determination may cause the Mach-band effect when we handle not a pixel but some pixels.

In order to solve the first problem, checking (recalculation) process is executed, if the primary approximation shows
the same as the reference pixel. For solving the second one, the Gouraud shading takes the place of the arithmetic mean of color.

2.2. Determination of Initial Rays to be Traced

In this subsection, a proposal is briefly explained. Suppose that size of an expected image is $M \times N$ pixels. For recursive determination whether a set of pixels is equated with the center pixel of the set, let $n$ be the smallest integer such that $(2^n + 1)$ is equal or greater than both of $M$ and $N$. In the following, we discuss image generation of a screen of $(2^n + 1) \times (2^n + 1)$ pixels. After the generation, the objective image with $M \times N$ pixels is taken from the expanded screen.

![Figure 1: Distributed pixels in case of $n = 3$. The previous method, such distributed rays are defined within a pixel having the initial ray. In the proposed method, there are some pixels far from the initial one.](image1)

In the previous method, distributed rays are defined within a pixel having the initial ray. In the proposed method, there are some pixels far from the initial one. Figure 1 shows a case of $n = 3$. Then a set of pixels which forms a subscreen with $(2^n + 1) \times (2^n + 1)$ pixels is determined by the similar procedure to the adaptive distributed ray tracing whether the pixels is equated with the center pixel or they should be separated. If the set is determined not to be equated, let $n$ be $n - 1$ and execute the same procedure recursively. In a case of $n = 1$ (see Fig.2), conventional ray tracing, i.e., distribution within a pixel, is applied to each pixel.

When the set is determined to be equated, color (brightness) of the pixels within the set are interpolated by the Gouraud shading from the center pixel and two primary approximated pixels. Figure 3 shows left-top area around an initial ray.

![Figure 2: The smallest region to be considered ($n = 1$). In such case, distribution of rays within a pixel is applied.](image2)

![Figure 3: Three distributed rays and an initial ray in the left-top area. If the results are different, subdivision is required. If not, color approximation over the area is necessary.](image3)

If the three distributed rays have different results from the initial one in the area given by Fig.3, then Gouraud shading should be applied to two area; a triangle having a pixel radiating the initial ray, the left and the left-top pixel as vertices, and a triangle having a pixel of the initial ray, the left-top and the top pixel as vertices (see Fig.4).

We note here that the set should be subdivided even if the set is determined to be equated, when the set is relatively large.

2.3. Algorithm

From the above, the proposed algorithm shown is as following.

Step 1: Define an expanded screen with $(2^n + 1) \times (2^n + 1)$ pixels according to a given screen.

Step 2: Set the $(2^n + 1) \times (2^n + 1)$ pixels as a target area, and a center pixel of the target area as a target pixel.

Step 3: Trace a ray, through the target pixel from a given viewpoint with automatic differentiation, i.e., calculate intersections, shadows, illuminance, and so on.
Step 4: Calculate the primary approximation at 8 pixels on the edge of the target area by the calculated with automatic differentiation. For each quarter of the target area, determine whether the three pixels in the same quartered target area can be equated with the target pixel in meaning of intersection with same object and the similar illumination with shadow. If it can be equated, go to Step 6. Else go to Step 5.

Step 5: Subdivide the target area in quarter with \( n = n - 1 \). If \( n = 1 \) go to Step 7. Otherwise return to Step 2.

Step 6: Using Gouraud shading, calculate color of all pixels in quartered target area.

Step 7: The target area is set to 3 \( \times \) 3 pixels. In this case, execute the adaptive distributed ray tracing on each pixel in target area.

The above proposed algorithm can reduce initial rays from a given viewpoint, especially when a scene includes wide empty space, background and/or flat surfaces on which the primary approximation works very well. As a result, the proposed algorithm contributes to less computation time with the similar quality under a certain scene satisfying the above.

### 3. Simulation

In this section, images with size of 513 \( \times \) 513 pixels are generated under the 750MHz UltraSPARC III system. In this case, firstly the target area (i.e. the expanded screen) is set to 513 \( \times \) 513 pixels (\( n = 9 \)). For making evaluation simply, we consider a sphere in a scene for an example.

Figure 5 shows an image generated by the proposed method. For reference, Fig.6 is generated by previously adaptive distributed ray tracing (Makino/Shinotsuka’s method), and Fig.7 is done by the conventional distributed ray tracing (Cook’s method). Also Table 3 shows computation time, number of initial rays and number of distributed (additionally defined) rays.

Comparing three figures, we can see that the proposed method gives lower quality than other two methods at the shaded area on the sphere (top-left region). In the area the edge of the shade seems stepwise in Fig.5, while it seems slantwise in Figs.6 and 7. Since the proposed method method skips exact calculation when the primary approximation far from a pixel is not different, such situation may be happened.

However, the proposed method works much faster than the two methods (20 times faster than Makino/Shinotsuka’s method, and 79 times faster than the conventional distributed ray tracing). Since a lot of unnecessary tracing and distribution of initial rays are prevented, this faster computation is achieved. On the other hand, Fig.6 is generated under tracing every initial ray through defined pixels on a screen, so that unnecessarily precise resolution and slower computation is achieved namely on dark area(background). Furthermore in Fig.7, all initial rays are distributed at the beginning of tracing.

Regardless of less traced rays, the proposed method works at edge of the sphere as well as other methods, since the same level of distributed ray tracing is executed at such area. Therefore, the proposed method is useful when both of fast computation and exact shape representation have priority over image quality on surface of an object. Expanding the proposed method to reflected, refracted and shadow rays, we expect to improve the quality.

### 4. Conclusion

In this paper we proposed an adaptive determination with automatic differentiation, from which pixels initial rays should be radiated. The proposed method can be applied anti-aliasing and benefits more efficient calculation of ray tracing. Improving quality of images will be discussed in future paper.
Acknowledgments

This paper is partially supported by “The Research on Security and Reliability in Electronic Society,” Chuo University 21st Century COE Program, Chuo University Grant for Special Research, and Collaborative Research Grant of the Institute of Science and Engineering, Chuo University.

References


Figure 5: An image by the proposed method with Gouraud shading. Outline is smoothly represented while shaded area seems stepwise. However, it works much faster than the others, since only 430 pixels are distributed among 599 initial rays on 263,169 pixels.

Figure 6: An image by the Makino/Shinotsuka’s method. Outline and shaded area are smoothly represented, although it works over 20 times slower than the proposed method, since 617 pixels are distributed among 263,169 initial rays and pixels.

Figure 7: An image by conventional distributed ray tracing. Outline and shaded area are smoothly represented, too. However, it works about 80 times slower than the proposed method, since all of 263,169 pixels are distributed.
Noise-assisted quantization

Shin Mizutani†, Jun Muramatsu†, Kenichi Arai† and Peter Davis†

†NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
Email: shin@cslab.kecl.ntt.co.jp, pure@cslab.kecl.ntt.co.jp, ken@cslab.kecl.ntt.co.jp, davis@cslab.kecl.ntt.co.jp

Abstract—We show that quantization with suboptimal thresholds can exhibit noise-assisted quantization in which information between analog input and quantized output is maximized by adding noise. Not all suboptimal quantization thresholds show noise-assisted quantization. The condition for noise-assisted quantization depends on the relation between the signal distribution and suboptimal thresholds. It is important to know the correct signal probability distribution and to calculate the correct optimal quantization thresholds to minimize information loss. However, in actual situation, we often and carelessly use uniform distributed quantization thresholds in A/D conversion because of a lack of prior knowledge or for simplicity. From this consideration, we may often find noise-assisted quantization in real signal processing.

1. Introduction

Noise-induced effects in nonlinear systems have recently received considerable attention. In particular, stochastic resonance (SR) [1] has been studied in various systems. SR means that the resonance response of a noisy nonlinear system to a subthreshold signal can be optimized by adding noise.

Information theoretical approaches have been used to study SR of aperiodic signals, for binary signals [2, 3]. In these studies, bit error probability or mutual information is used to measure transmission between input and output. SR in a threshold system with a signal as input and noise is closely related to the signal detection problem [2] in classical engineering studies [4]. For example, here we consider that binary signal has two kinds of values, that is 0 or 1 and noise has continuous values. Signal detection determines that the signal’s value is 0 or 1 from noisy input. The detection technique gives the optimal decision method that the input includes the signal or not. Full knowledge about the signal and noise can be used to obtain the optimal detector by calculating the optimal threshold. The optimization criterion is the maximal correct probability for detection. This corresponds to the maximal mutual information between input and output binary signal when we regard signal detection as a binary communication channel. An optimal detector with an optimal threshold shows monotonic decay of information against noise intensity. On the other hand, it is recently known that suboptimal thresholds can result in noise-assisted detection [5]. The noise intensity can maximize information. Here we note that mutual information using suboptimal threshold never exceeds information using the optimal threshold as shown by the data processing inequality [6]. One of reasons why we must use the suboptimal threshold is that we do not have full knowledge about the signal and noise. We can not calculate the optimal threshold for detection by a part of knowledge. When the threshold is calculated by our estimation about a signal and noise, the threshold may become suboptimal.

Here we consider noise-assisted quantization as that information between analog input and quantized output is maximized by adding noise. Quantization in engineering can be regarded as multi-threshold system. Quantization converts analog input signal to discrete output values representing each partition of signal values by quantization thresholds. From previous engineering studies, we can see a technique that a small amount of noise is added to analog signal before quantization to reduce distortion between input signal and output discrete values. This is called dithering. SR as dithering in multi-threshold systems have been reported in Ref. [7, 8]. However, in these previous SR studies, they used periodic signals as an input and signal-to-noise ratio as the input-output measure following classical SR studies. In general quantization, the input signal has continuous arbitrary values and the output has discrete values. Similar to signal detection, measures to evaluate quantization using certain threshold sets are mean square error or mutual information between analog input and digital outputs when we use the same number of quantization partitions.

Stocks has used an summing array of one bit threshold device and has shown SR with suprathereshold signal [9]. Input signal is Gaussian noise and environmental noise is also Gaussian noise with different parameters. If we consider that each threshold is located on different values, we can consider this array as a kind of quantizer. However, all threshold values are fixed with the same and independent noise is added to each threshold device. He showed that two and more thresholds are needed to see SR using mutual informa-
Our research aim is to show the simple engineering example with noise-assisted sensing similar to SR and to analyze this to find the general condition for noise-assisted sensing. From these, we can find general engineering mechanism of noise-assisted sensing and consider the applications in engineering using environmental noise or internal fluctuations positively. This approach is against the previous engineering that reduces the noise. However, noise and fluctuations are unavoidable in real world. From this study, we also may find the meaning of information processing with SR in biological systems.

In this paper, we focus on a simple quantizer to study noise-assisted quantization. Here we use signal and noise with certain probability distributions. There are optimization criteria commonly used for determination of quantization thresholds in engineering. One example is a criterion of minimum square error [10]. We consider that optimal quantization thresholds are calculated from the criterion of the maximal mutual information between input and output. This corresponds to quantization thresholds realizing equal probability for all partitions. We describe this in next session in detail. We use information also to compare the results on signal transmission by threshold [3] and the array of one bit threshold [9]. As in signal detection, we can imagine that a lack of knowledge to calculate the optimal quantization thresholds may result in observing noise-assisted quantization. We note here that the difference with the signal detection situation concerns with the signal. In signal detection, we consider signal’s values discrete and it has binary values in typical case. In quantization, signal has continuous values and it is quantized to discrete values.

In the next section, we explain the quantization model and show the formulation of mutual information. We give the examples of noise-assisted quantization in the third section and condition for noise-assisted quantization in the third section. Finally we conclude the results.

2. Quantization model

Scalar quantization is certainly used in analog to digital conversion that is needed in the beginning of digital signal processing. However, we usually use quantization thresholds located equally because we do not know the probability distribution and dynamic range of source. We consider the effect of noise added to input signal as in Fig. 1. We normally consider that noise only degenerates quantization. However, we show that noise can assist quantization when we use suboptimal thresholds.

Consider signal $x$ of continuous value with distribution $p(x)$ quantized to output $y_i$ ($i = 1, \ldots, N$) of discrete value using quantization thresholds $x_i$ ($i = 1, \ldots, N - 1$). Input to the quantizer includes signal $x$ and noise $\xi$ of distribution $p_\xi(\xi)$ with standard deviation $\sigma$.

First we explain optimization criteria to determine quantization thresholds when partition number $N$ is fixed. There are a few optimization criteria, however, to compare with the previous results as Ref. [9], we used is a criterion of maximal information. Mutual information $I(x; y_i)$ between input and output is as follows [9, 11].

$$I(x; y_i) = I(\sigma) = H(y_i) - H(y_i | x)$$

$$= \sum_{i=1}^{N} \int_{-\infty}^{\infty} dx \{ p(y_i | x) p(x) \}$$

$$\times \log \frac{p(y_i | x)}{\int_{-\infty}^{\infty} p(y_i | x') p(x') dx'} \quad (1)$$

$$= \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} p(x) dx \log \frac{1}{\int_{x_{i-1}}^{x_i} p(x') dx'}$$

$$= - \sum_{i=1}^{N} p_i \log p_i = H(p_i). \quad (2)$$

We can show that information $I(x; y_i)$ equals to entropy $H(p_i)$. Here $p_i$ is probability when output is $y_i$. We set that $\sum_{i=1}^{N} p_i = 1$. When $p_i = \text{const}$, we maximize information $I$ as $\max_{p_i} H(p_i)$. This corresponds to quantization thresholds realizing equal probability for all partitions. Here we note that the value $y_i$ does not have any relation to this mutual information calculation because we only need $i$ to discriminate partitions. However, when we use error between signal value $x$ and output $y_i$ as the measure, error needs exact value $y_i$ to calculate. We set the optimal threshold values $x_i$ calculated from the signal distribution without noise using equal probability.

To calculate information by Eq. (1), $p(y_i | x)$ is calculated as follows.

$$p(y_i | x) = \int_{-\infty}^{\infty} p(y_i | y) p(y | x) dx$$

$$= \int_{-\infty}^{\infty} p(y_i | y) p_\xi(y - x) dx$$

\[ \text{Figure 1: Quantization with noise.} \]
\[
\begin{align*}
\mathbb{P}(\xi|y) &= \begin{cases} 
\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right), & \text{if } y_1 \leq y \leq y_2, \\
0, & \text{otherwise}.
\end{cases}
\end{align*}
\]

Here \( y \) is the continuous value that equals \( x + \xi \). To calculate Eq. (3), we use the following equation about \( p(y|y) \).

\[
\begin{align*}
p(y|y) &= \begin{cases} 
1, & \text{if } x_{i-1} \leq y \leq x_i, \\
0, & \text{otherwise}.
\end{cases}
\end{align*}
\]

We also use that \( p(y|x) \) becomes distribution \( p_\xi(y-x) \) of noise \( \xi \) as follows.

\[
\begin{align*}
p(y|x) &= \int_{-\infty}^{\infty} p(y|\xi)p(\xi|x)dy \\
&= \int_{-\infty}^{\infty} \delta(y-\xi)p_\xi(\xi-x)dx \\
&= p_\xi(y-x).
\end{align*}
\]

3. Examples of noise-assisted quantization

For simplicity, here we consider that \( p(x) \) is uniform distribution \([0,1] \) with mean \( m = 0 \) and \( N = 4 \). The optimal threshold values become \( x_i = 0.25i - 0.5 \) (\( i = 1, \ldots, N-1 \)). We use Gaussian distribution \( N(0, \sigma) \) as noise distribution \( p_\xi(\xi-x) \). With the optimal thresholds, noise intensity \( \sigma \) only decreases information \( I(\sigma) \) as in Fig. 2.

![Figure 2: Mutual information \( I(\sigma) \) using the optimal thresholds and suboptimal thresholds. Dotted line shows information using the optimal thresholds, and line and dashed line show information using the suboptimal thresholds.](image)

Next, we consider that the signal has false probability distribution \( p'(x) \) different from true probability distribution \( p(x) \). We can also calculate false quantization thresholds \( x'_i \) from the false distribution. These quantization thresholds \( x'_i \) become suboptimal when the false distribution does not correspond to the true one. As the false distribution, here we use uniform distribution \([0,1] \) with mean \( m = 1 \) and calculate \( x'_i = 0.5i + 0.5 \) (\( i = 1, \ldots, N-1 \)) with \( N = 4 \). False uniform distribution \( p'(x) \) is located on the upper side of true uniform distribution \( p(x) \). All signal values from the true uniform distribution without noise become subthreshold from a viewpoint of suboptimal thresholds \( x'_i \). By these suboptimal quantization thresholds, we can calculate mutual information \( I(\sigma) \) using true distribution \( p(x) \) as in Fig. 2. We can find that there exists a maximum at moderate noise intensity \( \sigma \) in mutual information. This is what we call noise-assisted quantization: the mutual information between analog input and quantized output has a maximum value at a non-zero noise intensity. We also can find that there exists a maximum for false uniform distribution \([0,1] \) with mean \( m = 3/4 \). We do not show here, but we can see noise-assisted quantization with the case that both variances of true and false distribution are different.

4. Condition for noise-assisted quantization

Moderate noise transforms true distribution \( p(x) \) to be similar to false distribution \( p'(x) \) without noise. The mutual information of a true distribution \( p(x) \) with noise evaluated using suboptimal thresholds \( x'_i \) can exhibit the increase with increasing noise intensity. When false uniform distribution \( p'(x) \) is located on the lower side of true uniform distribution \( p(x) \), all signal values from true uniform distribution without noise become suprathreshold of suboptimal thresholds \( x'_i \). In this case, we can also see noise-assisted quantization. From these results, we can find that subthreshold or suprathreshold is not important for noise-assisted quantization. Threshold-crossing by noise is most important to find noise-assisted quantization. Noise can cause threshold-crossing, and then threshold-crossing results in noise-assisted quantization.

Not all suboptimal thresholds show noise-assisted quantization. For simplicity, using uniform signal distribution, we explain the condition to find noise-assisted quantization. The condition depends on the relation between the true distribution and suboptimal thresholds calculated from the false distribution without noise. From the above consideration, we could find noise-assisted quantization when the signal from true distribution without noise is perfectly in subthreshold or suprathreshold. Subthreshold and suprathreshold mean that the true signal without noise does not cross any suboptimal thresholds calculated from the false signal distribution. The direction across suboptimal thresholds, that is subthreshold or suprathreshold, does not related to noise-assisted quantization. We also note that the size of the variance of the true and false distribution is not related to noise-assisted quantization in subthreshold and suprathreshold cases.
Noise-assisted quantization also occurs when the true input distribution completely belongs between two thresholds of suboptimal thresholds. Therefore, the variance of the false distribution is fairly larger than the variance of the true distribution. We call this situation intra-threshold here. We also find noise-assisted quantization in the intra-threshold case. Here we can generalize subthreshold and suprathreshold into intra-threshold as follows. When we define quantization thresholds $x_i$ ($i = 0, \ldots, N$) again such as $x_0 = -\infty$ and $x_N = \infty$, subthreshold case becomes a special intra-threshold case between $x_0' = -\infty$ and $x_N'$ and suprathreshold case also becomes a special intra-threshold case between $x'_{N-1}$ and $x_N' = \infty$.

This intra-threshold case means that noise can enlarge the dynamic range of signal to quantize. The dynamic range and quantization thresholds are prepared for assumed distribution in general. Even when true signal values are far from false quantization thresholds, noise pushes the true signal values to the false thresholds to quantize.

Here we note that we can chose any probability distribution function for the signal and noise for noise-assisted quantization. For simplicity, we explained the condition for noise-assisted quantization in terms of the location of true signal distribution between suboptimal thresholds, that is subthreshold or suprathreshold or intra-threshold, using isolated uniform distribution for signal. However, intra-threshold including subthreshold and suprathreshold is sufficient not but necessary condition for noise-assisted quantization. We also can find noise-assisted quantization even when the signal distribution crosses a threshold a little without noise. This small amount of threshold-crossing to keep noise-assisted quantization is dependent on the model parameters. We also can find noise-assisted quantization in the case of nonisolated any distribution, for example, Gaussian distribution. If the input signal has a Gaussian distribution, it has a small amount of threshold-crossing without noise.

5. Conclusions

We showed that suboptimal quantization can cause noise-assisted quantization. It is important to know the correct signal probability distribution and to calculate the correct optimal quantization thresholds from the distribution. However, in actual situation, we often and carelessly use uniform distributed quantization thresholds in A/D conversion because of a lack of prior knowledge. From this consideration, we may often find noise-assisted quantization in real signal processing.

Not all suboptimal thresholds show noise-assisted quantization. We could find the condition for noise-assisted quantization that the true signal distribution is located between two suboptimal thresholds calculated from a false distribution. In the situation in intra-threshold that we called here, including subthreshold and suprathreshold, noise could transform the true signal distribution to be similar to the false distribution that produced suboptimal thresholds. In this paper, we mentioned examples of noise-assisted quantization only using isolated uniform distribution and Gaussian distribution for the signal. However, any signal distribution using almost all distribution functions could show noise-assisted quantization.

References

A Study on the Method for Evaluation of Pitch Variation of Baby’s Cry

Hidenori Tanaka† and Tomomasa Nagashima†‡

† Satellite Venture Business Laboratory, Muroran Institute of Technology 27-1, Mizumoto-cho, Muroran-shi, Hokkaido 050-8585, Japan
‡Department of Computer Science and Systems Engineering, Muroran Institute of Technology 27-1, Mizumoto-cho, Muroran-shi, Hokkaido 050-8585, Japan

Email: tanaka@mail.svbl.muroran-it.ac.jp, nagasima@epsilon2.csse.muroran-it.ac.jp

1. Introduction

Recently, the child abuse has become a social problem in Japan. The cause comes down to failure of communication between parents and children. So, it becomes an important social issue to construct strategy to prevent the problem. Development of a system-technology supporting the communication between parent and child appears to be useful. For the communication between human and human, the voice is one of the most useful tool. Especially, for babies whose faculties of speech are not developed sufficiently, cry is a very important tool to communicate. In the baby’s cry, it has been confirmed that the existence of nonlinearities [1]. Therefore baby’s cries are important subjects in the field of nonlinear dynamics. In the communication through cries from babies to parents, it has often made issues how various mental states of baby are reflected in the acoustic characteristics of cry. But, what emotions parents have by listening to cries is also very important problem. In the field of psychology, there are several studies about this problem [2]. However, in the field of sound analysis, there have been few studies to investigate the listeners’ impression of voice. Additionally, most of previous studies have targeted at nonverbal voices such as cries, but verbal voices. Therefore, we have been made some studies focused on relation between listeners’ impressions and acoustic features of babies’ cry.

Mukai et al. suggested that the cries of baby who have a structural disease in the vicinity of throat give listener the unpleasant impression and this cry is one of the causes of child abuse [3,4]. This disease is called Ankyloglossia with Deviation of the Epiglottis and Larynx (ADEL). Previously, we proposed the extended Pitch Perturbation Quotient (ePPQ) to quantify “the instability of the fundamental periods” that had been also pointed out as a feature of cries of ADEL by Mukai et al., and we suggested that the ePPQ was a candidate to evaluate the listeners’ unpleasantness for cry [5]. Additionally, we also suggested that the higher ePPQ cry has, the stronger that cry gives impressions which are related to child abuse [6]. However, the relation between listeners’ impressions and acoustic features has not been clarified yet.

There is a problem to be solved. Although there are several patterns of pitch variation, the difference of these patterns was not considered on the ePPQ. To improve the quality of evaluation for the listening impressions of cry based on the acoustic features, it is necessary to clarify the differences of these patterns on listening impressions. As a result of our observation of pitch variations, we found at least 3 kinds of pitch variation patterns, 1) random variation, 2) gradual variation and 3) sudden variation. In particular, it is found that variation in 3) looks quite similar to harmonic doubling. As the main difference of these pitch variation patterns seems to come from the way of the change of periodicity of corresponding sound wave, nonlinear analysis would be useful to clarify the differences on the mechanism of each pitch variation patterns.

2. Algorithms for Calculation of ePPQ

The ePPQ is an index to quantify the instability of fundamental periods. This was made by extension of PPQ that had been proposed by Kikuchi and Kasuya [7]. Here, the algorithm for calculation of ePPQ is expounded.

An ePPQ is calculated by substitution of time series of fundamental periods into $p(n)$ in following equation (1) and (2). The time series of fundamental periods was obtained by calculating the fundamental periods at regular time intervals in the subject of sound data.

\[
ePPQ = \frac{1}{N - 2k} \sum_{i=k}^{N-k-1} A(i),
\]  

(1)
Equation (1) denotes that ePPQ is average of \( A(i) \). We label \( A(i) \) as "perturbation value". Equation (2) denotes that the perturbation value \( A(i) \) is an absolute value of difference between 1 and current fundamental period divided by the average of fundamental periods for 2\( k+1 \) points which center on current time point.

To calculate the ePPQ, first of all, the time series of fundamental period must be made. In this study, we use the method applied the Average Magnitude Difference Function (AMDF) [8] to extract the fundamental period.

In this method, to begin with, following \( r(i) \) was calculated as an AMDF for a part of waveform data immediately after the time point \( t \) at which fundamental period should be obtained and a part of waveform data after \( l \) points delay.

\[
 r(i) = \sum_{m=t}^{t+M-1} |x(m) - x(m + l)|, \tag{3}
\]

where \( x(m) \) denotes the signal data of the sound and \( M \) was the length of the compared data. Then, the threshold \( r_{th} \) was calculated from the minimum value \( r_{min} \) and average value \( r_{av} \) of \( r(i) \) by following equation,

\[
 r_{th} = r_{min} + (r_{av} - r_{min})/4. \tag{4}
\]

Finally, the fundamental period was calculated by multiplying sampling period by minimal \( l \) which made \( r(i) \) less than or equal to threshold \( r_{th} \). The time series of fundamental periods was obtained by calculating the fundamental periods at regular time intervals in the same way. The ePPQ was defined by substitution of time series of fundamental periods into \( p(n) \) in equation (2).

3. Applying the ePPQ to real cry

In this chapter, we show an example of calculating ePPQ from babies’ cry practically.

Here, a baby’s cries that had been disclosed on Mukai’s homepage (http://www02.so-net.ne.jp/~s-mukai/) were used. These were cries of a baby who had ADEL and cries after the operation to remedy ADEL. It had been suggested that the cries of baby having ADEL give listener the unpleasant. These cry data were uttered by pinching lightly at the foot of a 1-month-old boy. To record the cries, MINIDISC RECORDER (MDS-102, SONY) was used. Microphone was set 15 cm above mouth of the babies. The data were downloaded from homepage and resampled to 40 [kHz]. Then voiced parts were cut out from them by visual observation. Here, we used a voiced part which was cut out from cries before the operation (data 1) and a voiced part which was cut out from cries after the operation (data 2) to calculate ePPQ.

The parameters for calculating ePPQ were set as below.

When the time series of fundamental period was made, the fundamental periods were calculated at the time interval \( \Delta t \) was set to 20 points (i.e., 0.5 [ms]). In equation (1) and (2) of previous chapter, \( k \) was set to 1 and \( N \) was set to the length of time series of fundamental periods for whole target sound data. In equation (3), the compared data length \( M \) was set to 128 and delay \( l \) ranged from 40 to 280 (i.e., \( l = 40, 41, ..., 280 \)).

Fig.1 and Fig.2 show the waveform and the time series of fundamental periods of data 1 and data 2. In each figure, the upper graph shows the waveform, where the horizontal axis shows time [s] and the vertical axis shows amplitude. And the lower graph shows the time series of fundamental period, where the horizontal axis shows time [s] and the vertical axis shows fundamental period [ms]. As a result of calculating the ePPQ, ePPQ for data 1 was 8.09×10⁻² and ePPQ for data 2 was 2.49×10⁻². Consequently, the cry of baby who has ADEL, it was said that it’s unpleasant, has higher ePPQ than the cry remedied by operation.
4. Discussions

There are several patterns in the pitch variation of babies’ voices. Robb and Saxman observed instances of harmonic doubling, fundamental frequency (\(F_0\)) shift, and biphonation in young children’s non-cry vocalizations [9]. Michelsson defined the harmonic doubling as a parallel series of harmonics that have the same melody form and occur simultaneously with the \(F_0\) and its harmonics [10]. And the biphonation is defined as a double series of non-parallel \(F_0\)’s where one can be falling and the other simultaneously rising. The \(F_0\) shift is seen as a break in the \(F_0\), denoted by rapid up or down movements of \(F_0\). Although Robb and Saxman observed these patterns in non-cry vocalizations, it seems that there are also patterns in cry.

We made time series of fundamental period for many cries using AMDF. Then we observed pitch variations. As a result, following 3 kinds of pitch variation pattern were found.

1) Random variation
Periodicity could hardly be found by visual observation of waveform (Fig.3). As a result of pitch extraction by AMDF method, fundamental periods varied randomly. It is possible that the variation include noise, \(F_0\) shift and biphonation.

2) Gradual variation
The fundamental period varied gradually by extension or contraction of pitch waveform.

3) Sudden variation
As the result of that the pitch waveform changes alternately, the length of a couple of former pitch waveforms becomes a fundamental period (for example, changing from fig.4(a) to fig.4(b)). Thereby, although the waveform changes gradually, fundamental period is doubled suddenly when double former fundamental period is considered more reasonable as the renewed fundamental period than single by AMDF method. The length of a former waveform is not so changed. This pitch variation is quite similar to the variation occurring by harmonic doubling.

In ePPQ, difference of these patterns was not considered. And thus any pitch variations were reflected equally in ePPQ. Among 3 patterns, gradual variation is seemed to effect hardly on ePPQ because it varies slowly. On the other hand, random variation and sudden variation are seemed to effect strongly on ePPQ because they vary on a large scale in the short time. Additionally, while the random variation has little periodicity, the sudden variation saves periodicity because the length of a former pitch waveform does not so change. Therefore, it seems that the listening impressions were different between random variation and sudden variation. Thus, for evaluation of listener’s impressions by acoustic features, it was necessary that distinction of these pitch variations. It seems that the main difference of these pitch variation patterns come from the way of the change of periodicity of corresponding sound wave. Therefore, it would be useful that applying of nonlinear analysis to clarify the differences on the mechanism of each pitch variation patterns.

![Fig.3: An example of random variation.](image)

![Fig.4: An example of sudden variation.](image)

5. Conclusions

In this paper, we have made the discussions to improve the quality of evaluation for the listening impressions of cry based on the acoustic features. Therat we showed pitch variation patterns found in baby’s cry and made discussions for the method of evaluating the difference of these patterns.
As a result of our observation of pitch variations for baby’s cry, we found 3 kinds of pitch variation patterns, 1) random variation, 2) gradual variation and 3) sudden variation.

In future works, it will be necessary to clarify the differences of the mechanism for each pitch variation patterns. It might be interesting to apply nonlinear analysis for this problem the main difference of these pitch variation patterns seems to come from the way of the change of periodicity of corresponding sound wave. Thereby, it might be expected to develop appropriate indices which can distinguish the pitch variation patterns.

Acknowledgments

We would like to thank Dr. Susumu Mukai for providing us with the materials and valuable advice.

References

Nonlinear Parametric Estimation for Signals in Nonstationary Random Noise via Stationarization and Wigner Distribution

Hiroshi Ijima†, Akira Ohsumi†, and Satoshi Yamaguchi†

† Department of Mechanical and System Engineering, Kyoto Institute of Technology, Matsugasaki, Sakyo, Kyoto 606-8585, Japan
E-mail: {ijima, ohsumi} @kit.ac.jp

Abstract—In this paper, a novel method is proposed for the estimation of parameters of signals corrupted by nonstationary random noise. Key approach is to modify the nonstationary observation data to stationary white ones using the theory of evolutionary spectrum and to calculate the Wigner distribution which is the quadratic form transform of the observation data. Then the parameter estimation is performed using the maximum likelihood approach for the likelihood function constructed by realizations of the Wigner distribution. The method is tested by simulations to show its efficacy.

1. Introduction

Up to the present time, the detection and/or estimation of signals corrupted by random noise have been investigated by many researchers. It has been commonly assumed that observation data is stationary and often Gaussian. In practice, however, most of the data encountered are nonstationary [1]. In recent years, wavelets have been advocated to treat the nonstationary data, but wavelets are not useful to estimate the unknown parameters attributed to the signal to be detected. Instead, the introduction of the Wigner distribution (WD) is attractive in the signal detection. The WD is bilinear in the signal (or noisy observation in our situation) because the signal enters twice in its calculation [2]. From the authors’ experience, the WD itself is not so much effective to detect the signal from the additive random noise even if it is stationary.

Combining the WD with the idea of maximum likelihood (ML), the authors have proposed a useful method of joint detection and estimation for signals which are corrupted by stationary random noise and have unknown parameters such as time-delay, attenuation coefficient, Doppler shift, and so forth [3,4]. Of course, the estimation of such unknown parameters is a kind of nonlinear parameter estimation problems.

It is the purpose of this paper to give an effective method of estimating parameters attributed to the signals which are corrupted by nonstationary random noise.

2. Problem Statement

Let \( s_0(t) \) be the scalar (real) sinusoidal signal

\[
s_0(t) = g(t) \cos \phi(t),
\]

where \( g(t) \) is the envelope of signal and \( \phi(t) \) is the phase function. When this signal is emanated from a remote source or transmitted actively to a target and received by a receiver, it is modulated in frequency, delayed in time, attenuated, and contaminated by random noise. So, when the signal is received, its form is considered as

\[
s(t) = a g(t - D) \cos \phi(t - D; f_0)
\]

where \( a, D \) and \( f_0 \) are the attenuation coefficient, time-delay, and the parameter related to the modulated frequency, respectively. Of course, in general, parameters are unknown. In this paper, for simplicity assume that \( a \) is known. The parameter \( a \) is linearly related to the signal, and hence the procedure of its estimation can be decoupled [5]. In order to specify the dependence of the unknown parameters we write \( s_0(t) \) as \( s_0(t; f_0, D) \).

The observation data is obtained in the following manner:

\[
g(t) = a s_0(t; f_0, D) + n(t), \quad t \geq 0,
\]

where \( n(t) \) is the additive random noise. The form of signal \( s_0(t) \) may be sure and its duration is assumed to be local in time. Estimating parameters \( D \) and \( f_0 \) may be important especially in the field of the measurement, control, communication and so on. To reflect the practical situation, we assume that the noise \( n(t) \) is nonstationary and generated by the stochastic model,

\[
dn(t) = -\beta(t)n(t)dt + \alpha(t)dw(t), \quad n(0) = n_0,
\]

where \( w(t) \) is a (scalar) standard Wiener process; \( \alpha(t) \) and \( \beta(t) \) are slowly and smoothly varying positive but unknown functions; and \( n_0 \) is a Gaussian random variable with zero-mean. Since (unknown) coefficients \( \alpha(t) \) and \( \beta(t) \) are time-varying, the noise \( n(t) \) as well as the observation process \( g(t) \) becomes inevitably nonstationary.
Then, our purpose is to propose a method for estimating unknown parameters $f_0$ and $D$ from the nonstationary observation data $y(t)$. The approach taken in this paper is as follows:

(i) First, the unknown coefficient functions $\alpha(t)$ and $\beta(t)$ in the noise model (4) are estimated using observation data $\{y(t)\}$.

(ii) Then, using the estimates of $\alpha(t)$ and $\beta(t)$, the nonstationary observation data $\{y(t)\}$ is modified to the stationary ones. Similar procedure has attempted recently to the detection problem of signals corrupted by nonstationary noise [6].

(iii) Based on the stationarized observation data, the estimation of unknown parameters $f_0$ and $D$ is achieved via the Wigner distribution-based maximum likelihood estimation (WD-MLE) developed by the authors [3,4].

3. Stationarization of the Observation Data

3.1. Estimation of Unknown Coefficients

First, the unknown coefficient functions of time, $\alpha(t)$ and $\beta(t)$, in the noise model (4) are identified. To do this, recalling that the duration of the signal to be detected is very local, let us consider the signal-free case, neglecting the signal’s existence,

$$y(t) = n(t),$$

which has the stochastic differential [7],

$$dy(t) = dn(t) = -\beta(t)y(t)dt + \alpha(t)dw(t).$$

(6)

Since we have assumed that $\alpha(t)$ and $\beta(t)$ change slowly and smoothly, they are assumed to behave approximately like a constant in an interval $I_t$ around the current time $t$,

$$\alpha(t) = \alpha_t, \quad \beta(t) = \beta_t \quad \text{for } t \in I_t.$$  

(7)

Then, the power spectral density of the $n(t)$-process is approximately evaluated as

$$S_t(\lambda) = \frac{\alpha_t^2}{\lambda^2 + \beta_t^2}. $$

(8)

The suffix in the notation $S_t(\lambda)$ stands for the dependence on the current time $t$. In this sense, $S_t(\lambda)$ may be interpreted as the time-varying spectral density or evolutionary spectral density in the sense of Priestley [8].

With the help of Priestley’s method for the estimation of evolutionary (time-varying) spectral density, the density $S_t(\lambda)$ can be estimated from the observation data $\{y(t)\}$. Let it denote by $\hat{S}_t(\lambda)$.

From (8), we have the relation

$$\frac{1}{\hat{S}_t(\lambda)} = \left(\frac{1}{\lambda^2} \right) \frac{\alpha_t^2}{\lambda^2 + \beta_t^2}.$$  

(9)

The coefficients $1/\alpha_t^2$ and $\beta_t^2/\alpha_t^2$ are estimated by minimizing the square-error, $[1/\hat{S}_t(\lambda) - 1/S_t(\lambda)]^2$ with respect to these coefficients. However, only the first one is accepted as a least-squares estimate to obtain $\hat{\alpha}_t = \sqrt{\hat{\beta}_t^2}$ because the estimate for $\beta_t$ obtained by this can not be recommended from the viewpoint of accuracy. Noting that the energy of the noise process is given in the neighborhood of $t$ by

$$v(t) = \int_{-\infty}^{\infty} S_t(\lambda)d\lambda = \frac{\alpha_t^2}{\beta_t} \pi,$$  

(10)

so its estimate $\hat{v}_t$ is calculated as $\hat{v}_t = \int_{-\infty}^{\infty} \hat{S}_t(\lambda)d\lambda$. From this, we have the estimate $\hat{\beta}_t$ by

$$\hat{\beta}_t = \frac{\alpha_t^2}{\hat{v}_t} \pi.$$  

(11)

Hence, we have the estimate $\hat{\beta}_t$ by substituting $\hat{\alpha}_t$ and $\hat{v}_t$ into (11).

3.2. Stationarization

By assuming further the signal-free case, the observation process (6) can be approximated locally using the estimated coefficients as

$$dy(t) = -\hat{\beta}_t y(t)dt + \hat{\alpha}_t dw(t).$$

(12)

This is expressed in the discretized version as follows:

$$\delta y_t = -\hat{\beta}_t y_t \delta t + \hat{\alpha}_t \delta w_t,$$  

(13)

where $\delta y_t (= y(t + \delta t) - y(t))$ and $\delta w_t (= w(t + \delta t) - w(t))$ are small increments of $y(t)$ and $w(t)$, respectively. Dividing both sides by $\hat{\alpha}_t \delta t$ provided that $\hat{\alpha}_t \neq 0$, we have

$$\frac{\delta y_t + \hat{\beta}_t y_t \delta t}{\hat{\alpha}_t \delta t} = \frac{\delta w_t}{\delta t}.$$  

(14)

Here, it should be noted that the right-hand side of (14) can be regarded as a stationary white Gaussian noise sequence with zero-mean and unit power spectral density. Keeping this fact in mind, let us define for each $t$ a sequence $\hat{y}_t$ by

$$\hat{y}_t = \frac{\delta y_t + \hat{\beta}_t y_t \delta t}{\hat{\alpha}_t \delta t}. $$

(15)

Hence, the sequence $\hat{y}_t$ can be regarded as a (discrete-time) stationarized version (for signal-free case) of the observation process $y(t)$.

Same argument can be possible for the case when the signal exists. Indeed, the observation process (3) is expected in the stochastic differential form, using the estimated coefficients as

$$dy(t) = a\delta_0(t; f_0, D)dt + d\eta(t) = a\{\delta_0(t; f_0, D) + \hat{\beta}_t s_0(t; f_0, D)\}dt - \hat{\beta}_t y(t)dt + \hat{\alpha}_t dw(t).$$

(16)
From this we have the discretized version,
\[
\delta y_t + \hat{\beta}_t y_t \delta t = a \{ \dot{s}_0(t; f_0, D) + \hat{\beta}_t s_0(t; f_0, D) \} \delta t + \hat{\alpha}_t \delta w_t. \tag{17}
\]

Dividing both sides again by \( \hat{\alpha}_t \delta t \), we have the expression
\[
\hat{y}_t = a \hat{\dot{s}}_0(f_0, D) + \gamma_t,
\tag{18}
\]
where
\[
\hat{\dot{s}}_0(f_0, D) = \frac{1}{\hat{\alpha}_t} \{ \dot{s}_0(t; f_0, D) + \hat{\beta}_t s_0(t; f_0, D) \}
\tag{19}
\]
and \( \gamma_t = \delta w_t/\delta t \) is the white Gaussian noise sequence. The expression (18) is familiar to us in the problem of detecting signals in stationary random noise.

4. ML Estimation Using Wigner Distribution

Hereafter, write \( \hat{y}_t \), as \( \hat{y}_t \) for \( t = \ell \Delta t \) (small increment of the time partition). Then, given the observation data \( \{ \hat{y}_t \} \), or \( \{ \hat{y}_t \}_{t=0,1,2,\ldots} \), generated by the discrete-time process (18), let \( W_{\hat{y}}(\ell, k; D) \) be the discrete WD represented in the following form [2]:
\[
W_{\hat{y}}(\ell, k; f_0, D) = \sum_{m=0}^{L-1} \hat{y}_{t+m} \hat{y}_{t-m} \cos \left( \frac{4\pi m k}{L} \right). \tag{20}
\]
This can be expressed as the sum
\[
W_{\hat{y}}(\ell, k; f_0, D) = W_{\hat{y}\hat{y}}(\ell, k; f_0, D) + 2W_{\hat{y}\gamma}(\ell, k; f_0, D) + W_{\gamma\gamma}(\ell, k), \tag{21}
\]
where
\[
W_{\alpha\beta}(\ell, k; f_0, D) = \sum_{m=0}^{L-1} \alpha_{t+m} \beta_{t-m} \cos \left( \frac{4\pi m k}{L} \right)
\tag{22}
\]
\[\alpha, \beta = \hat{s} \text{ or } \gamma.\]

It should be noted that in (21) there appear two terms due to the random noise \( 2W_{\hat{y}\gamma}(\ell, k; f_0, D) \) and \( W_{\gamma\gamma}(\ell, k) \), and that these two interfere in the legitimate auto-component \( W_{\hat{y}\hat{y}}(\ell, k; f_0, D) \) to detect the signal by observing the spectrum of \( W_{\hat{y}}(\ell, k; f_0, D) \) over the time-frequency domain.

Let \( W_A \) be realizations of the random field \( W_{\hat{y}}(\ell, k; f_0, D) \) over the time-frequency region \( \Delta = [0, T] \times [\lambda_{\min}, \lambda_{\max}] \times [T, \lambda_{\min}, \lambda_{\max}] \) (fixed). Let \( p \{ W_A | H^1 \} \) be the likelihood function of \( W_A \), where the notation \( H^1 \) implies that \( W_A \) is generated by (21). This likelihood function can be given by the limit of the density \( p \{ W_{11}, W_{12}, \ldots, W_{LM} | H^1 \} \) for discretized realizations \( \{ W_{ij} \}_{i=1,2,\ldots, L; j=1,2,\ldots, M} \) over \( \Delta \). Let \( w \) be an \( LM \)-vector with elements \( \{ W_{ij} \} \) arranged in lexicographic order. Then, the likelihood function of \( w \) is expressed by a Gaussian function [3,4]:
\[
p \{ w | H^1 \} = (2\pi)^{-\frac{LM}{2}} | R_1 |^{-\frac{1}{2}} \cdot \exp \left\{ -\frac{1}{2} (w - m_1)^T R_1^{-1} (w - m_1) \right\}, \tag{23}
\]
where \( m_1 \), \( R_1 \) are mean and covariance matrix (under \( H^1 \)). For convenience’ sake, let \( p \{ w | H^1 \} \) be the likelihood function constructed under the hypothesis that \( W_{\hat{y}}(\ell, k; f_0, D) = W_{\hat{y}\gamma}(\ell, k) \). We compute the ratio of these likelihood functions \( p \{ w | H^1 \} / p \{ w | H^0 \} \) to estimate unknown parameters \( f_0 \) and \( D \). In practice, we employ its logarithmic form:
\[
L(w; f_0, D) = -\frac{1}{2} \left\{ \ln | R_1 | - \ln | R_0 | \right\}
\tag{24}
\] + (w − m_1)^T R_1^{-1} (w - m_1)
− (w − m_0)^T R_0^{-1} (w - m_0)\right\},
\]
where \( m_0 \) and \( R_0 \) are the mean and covariance under \( H^0 \). Hence, by maximizing this function with respect to \( f_0 \) and \( D \), we can obtain WD-MLEs \( \hat{f}_0 \) and \( \hat{D} \), i.e.,
\[
\hat{f}_0, \hat{D} = \arg \left\{ \max_{f_0, D} L(w; f_0, D) \right\}. \tag{25}
\]

5. Simulation Studies

To confirm the usefulness of the proposed method several simulation experiments were performed.

In the following simulations, the observation data is generated by the discretized versions for (2) and (3) and with the time step size \( \Delta t = 1 \) s.

The time-varying coefficients of the nonstationary noise process \( n(t) \) were set as \( \alpha(t) = 0.91 - 0.8 \cos(0.006t - 3.066) \) and \( \beta(t) = 0.5 - 0.5 \cos(0.006t - 4.293) \). A Gaussian-shaped chirp signal (1) with \( g(t) = 1.6 \exp(-t^2/1000) \) and \( \phi(t; f_0) = 2\pi(-0.0013 t^2 + f_0 t) \) is employed. The true parameters for \( D \) and \( f_0 \) were set as \( D^* = 800 \) s, \( f_0^* = 0.25 \) Hz, respectively. The top figure in Fig.1 depicts the observation data during the time interval \([0, 1000]\) s. The signal to be detected is illustrated on the middle. The bottom in Fig.1 depicts the modified (stationarized) observation data \( \hat{y}_t \). Though the middle part is smaller than the other parts, as a whole it can be said that the sequence \( \hat{y}_t \) is well stationarized compared with the original nonstationary process \( y(t) \).

Fig.2 depicts the likelihood-ratio function \( L(w; f_0, D) \) over the parameter domain. It is apparent from Fig.2 that \( L(w; f_0, D) \) takes a distinctively large peak around the point \( (f_0, D) = (0.25, 800) \) s, and this fact proves without doubt the existence of the signal. From this we can say that the proposed method has an excellent performance. The maximization of the likelihood function...
in (25) was performed for grids over a possible region
$[0, 0.5] \times [1, 1024]$ of $(f_0, D)$ with each grid width 0.02 and 1. The computation time was 30 s.

The estimated coefficients $\hat{\alpha}$ and $\hat{\beta}$ are illustrated in Fig.3. For the estimation interval $I_t$ was set as $I_t = 360$s. The dash-dot line indicates their true values. From Fig.3, we see that $\alpha(t)$ and $\beta(t)$ are well estimated except around the first peak of $\beta(t).

6. Conclusion

In this paper, a method of parameter estimation of signals from the observation data corrupted by nonstationary random noise was proposed based on the modification of nonstationary observation data to stationary ones. Then, the modified stationary data has been applied to Wigner distribution-based estimation method, and the efficiency of the proposed method has been confirmed by simulation studies.

References


An Iterative Warping Algorithm for Arbitrary Frequency Maps

Salvatore Caporale, Luca De Marchi and Nicolò Speciale

ARCES/DEIS University of Bologna
Viale Risorgimento 2, Bologna, Italy
Email: scaporale@arces.unibo.it

Abstract—In this work we present an algorithm to compute frequency warping according to arbitrary shaped maps. We extend the well–known method based on Laguerre transform with an approach whose computational structure consists on a recursive scheme. The algorithm complexity and performances are discussed, moreover new perspectives and potential applications are shown, which include the combination of the warping algorithm with the Wavelet Transform.

1. Introduction

In the last years time–frequency transformation techniques acquired a leading role in signal processing and were widely applied. However, such transformations have restrictive properties which make them not suitable in some applications, in particular, the possibility of generalizing and adaptively varying the time–frequency domain tiling is a major aim.

In order to accomplish this task many strategies are possible, such as the application of a preliminary invertible transformation (warping) to reshape the frequency axis in an invertible and flexible way [1]. The results of this pre-processing procedure are then transformed according to a specific time–frequency analysis, such as the Wavelet Transform, so that the latter transformation is modified in a proper way by the former. In addition, the frequency axis warping can be defined so that orthogonality is preserved.

Frequency warping has been introduced some years ago [2] and then considered as a projection on a set of frequency and amplitude modulated functions (FAM) [3]. It suffers two major drawbacks. The former is represented by the computational complexity, which is $O(N^2)$. In spite of the availability of a recursive algorithm, very efficient from an architectural point of view, complexity affects latency and makes real–time implementation very hard to manage. The latter is related to the commonly-adopted Laguerre transform approach [4, 5]. This frequency warping implementation accomplishes computational exactness but allows for a reduced set of mapping functions which can be used to reshape the frequency axis and vary its sampling for a given transformation.

In this work we propose a solution to this second limitation. We extend the approach based on Laguerre transform allowing the employment of any kind of frequency axis re-shaping function rather than a specific class of functions. This generalization is accomplished without modifying the iterative algorithm basic structure and without significantly increase of computational complexity. Finally, though the transformation exactness is compromised, however an arbitrary precision can be easily reached.

2. Frequency Warping in discrete-time spaces

Let us suppose that we want a deformation of $x$ axis with a proper warping function $w(x)$. In order to guarantee invertibility, we have to choose $w(x)$ so that it maps $x$ axis on itself, that is:

$$\frac{dw}{dx} > 0 \ a.e. \ \Rightarrow \ \exists w^{-1}, w^{-1}(w(x)) = 1.$$ (1)

To preserve orthogonality, the axis warping has to be a linear, non singular and isometric transformation, that is it must preserve distance. Given a generic function $s(x)$, the warping operator $W$ is defined as:

$$[Wa](x) = \sqrt{\frac{dw}{dx}} s(w(x))$$ (2)

where linearity is immediate, the surjective property is guaranteed by (1) and isometry is easy to demonstrate:

$$\|Ws\| = \int_R w'(x) s(w(x))^2 dx = \int_R |s(y)|^2 dy = \|s\|.$$ (3)

So the warping operator results in a unitary transformation which preserves orthogonality, whose inverse is directly obtained by transposing and conjugating $W$.

To find the integration kernel of the warping operator, we first define it for a generic function in $L^2(\mathbb{R})$, then the definition is specified for frequency axis warping and finally extended for a discrete-time function in $\ell^2(\mathbb{Z})$.

The kernel of (2) could be easily defined as follows:

$$K_W(x,y) = \sqrt{\frac{dw}{dx}} \delta(w(x) - y)$$ (4)

but this definition is almost useless from a computational point of view. By using the Fourier domain we can achieve an alternative expression of integration kernel $K_W$:

$$K_W(x,y) = \sqrt{\frac{dw}{dx}} \int_{\mathbb{R}} e^{-j2\pi z(w(x)-y)} dz$$ (5)
which can be written as the composition of an inverse Fourier transform $F$ and a modified Fourier transform $F_w$, still unitary, defined as follows:

$$[F_w s](f) = \sqrt{\frac{dw}{df}} \int_{\mathbb{R}} s(t)e^{-j2\pi \omega(f)t} \, dt$$  \hspace{1cm} (6)

where $w(f)$ is always an odd function. By doing so, the warping operator is compactly expressed by:

$$W = F_w F^{-1}.$$  \hspace{1cm} (7)

Consequently, the passage from axis warping $W$ to frequency axis warping $W_f$ is expressed by:

$$W_f = F^{-1} F_w F^{-1} F = F^{-1} F_w.$$  \hspace{1cm} (8)

Now let us consider $s$ as a discrete-time function. Since we have illustrated the frequency axis warping in terms of operators, the passage to discrete-time domain is almost trivial: the operator $F$ becomes the ordinary Discrete-Time Fourier Transform, while we have to redefine $F_w$ and its inverse $F_w^{-1}$:

$$[F_w s](f) = \sqrt{\frac{dw}{df}} \sum_{n \in \mathbb{Z}} s(n)e^{-j2\pi \omega(f)n}$$  \hspace{1cm} (9)

$$[F_w^{-1} s](n) = \int_{0}^{1} \sqrt{\frac{dw}{df}} \hat{s}(f)e^{j2\pi \omega(f)f} \, df.$$  \hspace{1cm} (10)

where $\hat{s}$ denotes the Fourier transform of $s$, $w(f)$ is defined in the interval $[-1/2, 1/2]$ and extended as $w(f + k) = k + w(f)$, with $k \in \mathbb{Z}$.

Equations (8) and (10) completely define the frequency warping operator for discrete-time signals, which is actually a matrix of infinite dimensions. In real-life problems we need to limit the input sequence length to $N$ samples and find a way to limit output sequence length to $N'$ samples. Since $W$ must be invertible, $N' \geq N$. In next section we will show that, although in theory $N'$ is infinite, it can be limited without significant loss in precision. Since we are considering no input sequences in $L^2(\mathbb{Z})$, warping is no more a unitary transformation, rather a redundant transformation, whose matrix has $N$ orthogonal columns, but $N'$ non orthogonal rows.

3. Generalized iterative algorithm

With reference to the inverse function $w^{-1}$, another expression for frequency warping operator $W_f$ in terms of the operators $F$ and $F_w^{-1}$ could be defined:

$$W_f^{-1} = F^{-1} F_w^{-1} \Rightarrow W_f = F_w^{-1} F$$  \hspace{1cm} (11)

The kernel of $W_f$ is expressed by:

$$K_{W_f}(k, n) = \int_{0}^{1} \sqrt{\frac{dw^{-1}}{df}} e^{j2\pi \omega^{-1}(f)\cdot f} \, df.$$  \hspace{1cm} (12)

Figure 1: Scheme of the iterative frequency warping algorithm. The main difference between the generalized algorithm and the Laguerre case is the presence of the upper branch which calculates advance samples. In general, the filters are not causal.

Let us consider $k \in \mathbb{Z}$ as a parameter and let refer to rows of $W_f$ as $h_k(n)$. These could be seen as filter impulse responses, so that their transfer functions are $H_k(f)$. In terms of operator $H_k(f)$ can be obtained by applying $F$ to rows of $W_f$, that is $F$ must be applied right of $W_f$:

$$W_f = F^{-1} \cdot FF$$  \hspace{1cm} (13)

where $FF$ is simply a sign inversion. It follows:

$$H_k(f) = \sqrt{\frac{dw^{-1}}{df}} e^{j2\pi \omega^{-1}(f)} = \sqrt{\frac{dw}{df}} e^{j2\pi \omega^{-1}(f)}.$$  \hspace{1cm} (14)

By defining the pre-emphasis filter $H_0$ and the iteration filter $H(f)$ as:

$$H_0(f) = \sqrt{\frac{dw^{-1}}{df}}$$  \hspace{1cm} (15)

$$H(f) = e^{j2\pi \omega^{-1}(f)}$$  \hspace{1cm} (16)

respectively, $H_k(f)$ can be represented in a recursive way:

$$H_k(f) = H_{k-1}(f)H(f) \quad k \in \mathbb{Z}_+$$  \hspace{1cm} (17)

$$H_k(f) = H_{k+1}(f)H(-f) \quad k \in \mathbb{Z}_-.$$  \hspace{1cm} (18)

Moreover, by defining $\tilde{s}(n) = s(-n)$, we can write:

$$s_n(k) = \langle k, h_k \rangle = \sum_{n=0}^{N-1} K_{W_f}(k, n) s(n) = \sum_{m=n}^{n+N-1} h_k(m) s(n-m) \bigg|_{l=0}^{n+N-1}.$$  \hspace{1cm} (19)

Finally equations (17) and (18) can be used in order to formulate a recursive algorithm for frequency warping:

$$s_0 = h_0 \ast \tilde{s}$$  \hspace{1cm} (20)

$$s_k = h \ast s_{k-1} \quad k \in \mathbb{Z}_+$$  \hspace{1cm} (21)

$$s_k = \tilde{h} \ast s_{k+1} \quad k \in \mathbb{Z}_-.$$  \hspace{1cm} (22)

where $h(n) = h(-n)$.
The computational structure is schematically depicted in figure 1. The warped signal $s_w$ is obtained by sampling the correlation sequences $s_k$:

$$s_w(k) = \langle s, h_k \rangle = s_k(0) \quad \forall k \in \mathbb{Z}.$$  \hfill (23)

Except in the Laguerre case, every warping function $w(f)$ generates non-causal pre-emphasis and iteration filters, as shown in equations (18) and (22). However, the presence of advance samples is not a major issue as far as frequency warping produces additional delay samples, due to intrinsic disalignment of frequency components. It is worth to note that $w(f)$ can be chosen arbitrarily, so functions with flexes can be considered as well.

The number of output samples in equation (23) has to be limited in the same way as in the Laguerre case. The frequency warping is not a stationary transformation, then the indexing of input sequence affects output. If the input sequence has $N_p$ positive indexed samples and $N_n$ negative indexed ones, the number of non-negligible samples ($N'_p$ and $N'_n$) in the output sequence could be estimated by the maximum of the derivative $\dot{w}$ of the warping map:

$$N'_p \sim N_p \max \dot{w}$$ \hfill (24)

$$N'_n \sim N_n \max \dot{w}$$ \hfill (25)

although the number of samples to be computed could be a little increased in order to guarantee a given reconstruction accuracy. A schematic representation of a generic warping matrix is given in figure 2. In the case of warping function $w(f)$ with flexes, this approximation is not accurate and the number of output samples can be empirically obtained.

Figure 2: Structure of a generic warping matrix for an input signal indexed in $[-N_n, N_n]$ with $N_n = N_p$. The element of indexes $(0,0)$ acts as a center of punctual symmetry, that is $K_{W_p}(k,n) = K_{W}(-n,-k)$.

$$-N_n, 0, N_p$$

$$-N_n \max \dot{w}, 0, N_p \max \dot{w}$$

$$-N_n \min \dot{w}, 0, N_p \min \dot{w}$$

$$\sim 0$$

$$-N_n, 0, N_p$$

Figure 3: Basic scheme for the implementation of non-causal filtering. The state of IIR filters must be reset when the output buffer is full, since time-reversing the entire infinite response would produce temporally non-ordered samples. Resetting the state is also necessary in order to iteratively use the filter.

4. Implementation details and performances

The non-causal filters $h_0(n)$ and $h(n)$ can be implemented by a parallel decomposition, splitting them in causal and anticausal parts, calculating separately the two responses and then summing them, as schematically illustrated in figure 3. This can be achieved by adding at $k$-iteration step latency to realize time reversing of correlation sequence $s_k$. Obviously, at each iteration step the correlation sequences must be truncated, but this operation does not affect significantly the accuracy. We obtained good performances by storing at each step $4N$ samples, so that the delay of the non-causal filter results $8N$ and total delay is about $8N^2 \max \dot{w}$. Unlike the Laguerre case, only the iterative implementation is possible and the delay can not be reduced to $N$ by serializing the filters chain.

The causal and anticausal filters must be approximated: an approximation based on IIR filters allows a good accuracy with a filter order proportional to $\log N$, so complexity results $\propto N^2 \log N$.

Since the pre–emphasis filter $h_0$ is symmetric, the causal ($h_c^0$) and anticausal ($h_a^0$) filters are equal:

$$h_c^0(n) = h_a^0(n) = \frac{1}{2} h_0(0) \delta(n) + h_0(n) u(n-1)$$ \hfill (26)

where $u(n)$ is the unitary step sequence. As regards the iterative filter $h$, which is not symmetric, the splitting is made so that the causal part $h'^c$ and the anticausal part $h'^a$ result to be minimum phase:

$$h'^c(n) = h(n+1) u(n)$$ \hfill (27)

$$h'^a(n) = h(-n) u(n).$$ \hfill (28)

As a consequence equations (20)-(22) become:

$$s_0(n) = [h_0^0 + s](n) = [h_0^0 + \delta](n)$$ \hfill (29)

$$s_k(n) = [h^c * s_{k-1}](n-1) + [h^c * s_{k-1}](n-1) \quad k \in \mathbb{Z}_+$$ \hfill (30)

$$s_k(n) = [h^a * s_{k+1}](n+1) + [h^a * s_{k+1}](n) \quad k \in \mathbb{Z}_-.$$ \hfill (31)

The truncation of correlation sequences $s_k$ can be made through a static or a dynamic window.

By using $K \log_2 N$-order filters, we obtain a signal to reconstruction error ratio of about $90 \div 100$ dB, which is sufficient for most applications.
5. Applications

Thanks to flexibility on the choice of the warping function \( w(f) \), generalized frequency warping can be applied to a lot of problems. The most immediate application is the modification of wavelet analysis properties. Let us suppose we want to obtain a multiresolution analysis which characterizes frequency axis by an exponentially decreasing resolution. This is the case of Discrete Wavelet Transform, but it achieves a poor frequency resolution, since it splits the input signal in sequences representative of octaves. The application of a Wavelet Packet analysis on each octave increases frequency resolution analysis, but on the other hand it results to be constant on each octave. So a frequency warping can be usefully interposed between DWT and WP. To have an exponentially decreasing resolution for the positive frequency axis we use the following warping function:

\[
w(f) = \frac{1}{2} (4^f - 1) \quad f \in [0, 1/2]
\]  

(32)

depicted in figure 4, whose warping matrix is shown in figure 5. It is worth to notice that this kind of warping can be achieved only by the generalized algorithm introduced in the previous section. Unlike Laguerre case, our approach allows to use warping functions with flexes.

6. Conclusion

In this work we proposed an algorithm which allows for the calculation of warping transformations with arbitrary frequency maps. The computational structure is based on the iterative application of IIR filters. We discussed the algorithm complexity and performance, and we suggested how the non-stationary warping procedure could be combined with the Wavelet Transform to have a flexible tiling of the time–frequency plane.

References


T-entropy of EEG/EOG Sensitive to Sleep State

M.R. Titchener†

†Department of Computer Science, University of Auckland
Private Bag, Auckland, New Zealand
Email: mark@tcode.auckland.ac.nz

Abstract—The use of symbolic measures of complexity to derive quantitative characterisations of temporal structures in time series is proving potent for a variety of areas. Such techniques have particular attraction in the medical field, where clinical diagnosis and/or application of therapies is reliant on electronic patient monitoring systems. This paper explores the application of a relative new comer to the area of non-linear measures, using T-entropy computed from electroencephalogram (EEG) and electrooculogram (EOG) signals to derive indicators of sleep state.

1. Introduction

A growing body of literature is devoted to applying non-linear measures to medical and biological time series [1, 2]. One area that has attracted particular attention involves quantifying anesthetic depth from EEG/EOG signals [3, 4]. Many of the recognised measures are statistical in nature. Some are purely statistical. Others involve coarse grained encoding of time series data to obtain symbolic representations of the dynamics. The latter methods are predominantly couched in probabilistic terms, reflecting Shannon’s information theoretic notion of entropy.

A further class of measures derive from the study of non-linear systems, and symbolic dynamics. The discovery of the duality that exists between deterministic and stochastic descriptions for certain dynamical systems almost certainly inspired Kolmogorov’s ideas on algorithmic entropy. Contrasting with Shannon’s probabilistic formulations, Kolmogorov felt it was possible to define information meaningfully for individual finite strings. Later called the (Chaitin-)Kolmogorov complexity, and subsequently proven uncomputable, Kolmogorov’s algorithmic definition led Lempel and Ziv [10] to a more practical definition of string production complexity. Watanabe et al [9], and others have applied the LZ measure to the analysis of biological/medical and other data series with some success.

This paper uses alternative measures, the T-complexity and its linearised counterparts, T-information and T-entropy, in analysis of EEG/EOG time series. The aim of this work has been exploratory, to identify whether known sleep states are discernable from these measures. In ignorance of the work of Lempel Ziv, the T-complexity definition was proposed by the author in 1993, following in the algorithmic footsteps of Kolmogorov and Chaitin. T-complexity may be viewed as belonging to a general class of string production measures such as Lempel and Ziv considered. However, significant differences in the respective parsing algorithms result in contrasting complexity definitions and in the case of the T-complexity, computational advantages including linearisation of the measure.

Lempel and Ziv defined their string production complexity to be the size of the smallest vocabulary of patterns necessary to reproduce the string from its alphabet. Whereas the LZ76 complexity measure derives from the application of an exhaustive linear-pattern copying search algorithm, computing the T-complexity involves a parsing algorithm that results in a unique recursive hierarchical-pattern copying prescription for the string. While the two measures exhibit similar sensitivities to detecting pattern structure, establishing a formal link between the two measures remains an open problem. One striking difference between the two measures is in their respective computational speeds; $O(n)$ for the T-complexity, and $O(n^2)$ for the LZ76 algorithm (based on using fixed word-size machines). While variants of the LZ76 algorithm, the LZ77 and LZ78 specifically tackle the matter of computational effort, these do so in ways that ultimately compromise the measurement of string complexity.

An upper bound for the T-complexity turns out to be quite precisely given by a logarithmic integral function, whose inverse is used to linearise the measure in relation to string length. In this paper we are specifically interested in using the linearised measure, T-entropy to explore EEG/EOG patient data. Combining the respective T-entropy series in a ‘pseudo’ phase-space reconstruction ultimately leads us to concluded that T-entropy may be used to give a useful indicator of state.

2. Symbolic dynamics: encoding time-series

Symbolic measures involve translating real-valued time-series into symbolic strings through coarse grained partitioning of the ‘state space’. The choice of partition generally revolves around maximising the resultant string entropy. The supremum of entropies computed over all possible partitions (finite or infinite) yields the Kolmogorov-Sinai entropy (KS-entropy), a characteristic property of the dynamics.

In practice, time-series data represents a projection of a considerably reduced subset of a d-dimensional phase
space. Such a series is likely to be distorted by noise effects and other systematic errors. Coarse grained encoding of a time series has the advantage of reducing sensitivity to noise without necessarily sacrificing the essence of the dynamics. Computational requirements for processing the discrete symbolic representations is enhanced. Dynamical systems like the logistic map have often been used to evaluate the effectiveness of symbolic measures over the range of possible chaotic behaviours. [6] shows the T-entropy measure to be effective in mirroring the positive Lyapunov exponent, and by implication, the KS-entropy for the logistic map, over the whole of the chaotic regime. Its sensitivity to the range of dynamics suggests that it will perform usefully on, for example, medically derived data series.

3. Deterministic T-entropy

The definition of T-complexity hinges on the observation [5] that given an alphabet \( A \), any string \( x \in A^n \) may be expressed as a catenation of recursively derived patterns, that is:

\[
x = p_q^k q_{q-1} \ldots p_1^k a,
\]

\( k_i \in \mathbb{N}^+ \), \( a \in A \), and \( p_i \in p \) where

\[
p = \{ p_i | p_i = p_{i-1}^{m_{i-1}} \ldots p_{j}^{m_{j}} \ldots p_1^{m_1}, a_j, a_i \in A, p_1 \in A \text{ and } m_{j,i} \in \mathbb{N} \}
\]

The process by which a given string is decomposed to patterns \( p = (p_1 \ldots p_q) \) and copy factors \( k = (k_1, \ldots, k_q) \), consistent with Eqns (1)& (2) is referred to as T-decomposition. The T-complexity, \( C_T(x) \), of a string \( x \) is defined as a weighted count of the number of patterns:

\[
C_T \equiv \sum_{i=1}^{q} \log_2(k_i + 1).
\]

If a single pattern occurs for each pattern, the complexity simply reduces to \( C_T = q \), the size of the pattern vocabulary. Hence the T-complexity is essentially a weighted measure of the number of steps required to produce the string, and is thus ascribed units of taugs (T-augmentation steps).

Decomposition of a string comprising \( n \) symbols, to its constituent \( q \) patterns is achievable at speeds of \( O(n) \). This efficiency lends the measure to medical or industrial signal processing applications where real-time (or near real-time) performance is required.

The T-complexity upper-bound [7] is \( b_T(n) \approx li(n \ln \#A) \) where \( \#A \) is the size of the alphabet and \( li() \) is the logarithmic integral function. The factor, \( ln \#A \), may be interpreted as implying conversion of units. The string length, counted in alphabet symbols, is thus converted to a length expressed in nats, a measure of information that presumes the natural logarithm rather than the logarithm to base 2.
The T-information is the linearised measure:

$$I_T \equiv \text{li}^{-1}(C_T \text{ (taugs)}) \text{ (nats)}.$$  

The average T-information rate, i.e., information per symbol referred to as the T-entropy, is:

$$H_T \equiv \frac{I_T}{n} \text{ (nats/symbol)}.$$ 

Software tools have been developed to efficiently compute and display the computed T-complexity, -information and -entropy values from individual strings, or from partitioned time series data. In the case of the latter user defined parameters determine the dynamic range and window resolution generating data arrays that may be visualised as three-dimensional surfaces to expose structural features of potential interest.

4. T-entropy of EEG/EOG time-series

Figure 2 depicts entropy surfaces for EEG/EOG data series comprising just over a million and half a million sample points respectively, with just 21 bi-partition levels (c.f. Fig. 1 method B). Using a current hi-end laptop, the time to produce these surfaces is less than 10 seconds. Here, the window sizes are 6000 & 3000 points for the EEG & EOG signals (sampled at 100 and 50Hz respectively), a 50% overlap of window positions provides a degree of smoothness, and results in an array of approximately 760 × 21 ≈ 1,500 vertices for each surface. Surface smoothing is further provided in the graphics display software [11] which also allows flexible 3D viewing and lighting of the graphical objects.

To date, the T-entropy for some 45 patient data sets\(^1\) from which a number of common observations have been identified. These are illustrated in relation to the above given example.

The breadth of each surface derives from the swept partition, and surface topology corresponds to the magnitude of the T-entropy computed from sample strings reflecting the partition levels and window positions accordingly. Each surface individually depicts the sleep activity for a whole nights data and allows one to readily identify regions of interest. Three example areas of interest are highlighted with arrows. Using manually scored sleep states it has been possible to identify these as corresponding to known S2/S3/S4 sleep states.

The “mountainous” cross section implies that the key

\(^1\)From Prof. D. Rapoport of the New York Medical Center, as de-identititied files obtained after IRB approval of waiver of consent.
dynamics are effectively captured by the maximal-entropy profiles. The scaled up profiles superposed as red and green plots may be compared directly with one another since they share common units (milli-nats/sample), time scale and resolution. Figure 3 (top) presents two such graphs plotted together for three separate patient data sets. What is interesting here is that the areas of initial interest in Figure 2 appear very much as universal feature in the sleep studies looked at to date. These are identifiable by the green curve rising above the red, followed by a period of downward drift, mirrored simultaneously in both profiles.

To make further sense of the entropy functions these are graphed one against the other as a function of time, creating a ‘trajectory’ (c.f. Figure 4, top). A series of linear partitionings of the H(C3)-H(LOC) entropy plane divide this into regions which strongly correlate with manually scored states derived for these sleep studies, and reveal new perspective (c.f. Figure 4, bottom). The regions are here colour coded to reflect the partitioned areas. Projecting the points into the plane we obtain a scattergram, whose coded regions found to be broadly consistent across the patient data, although not all patients exploit the full range of sleep states in a given night’s study. In many cases states may appear only fleetingly. Nevertheless, this simplistic partitioning using entropy to classify the states shows good correspondence with the manually scored sleep states (c.f. Figure 5) derived according the rules set out in [8].

5. Concluding remarks

The computed T-entropy for EEG/EOG time series reveals useful sensitivity to known sleep states. However the results have so far come from patients identified as suffering from sleep disorders. It will be evident that considerable scope exists for extending this approach for analysis to more widely representative groups, to establish the norms and to further identify if T-entropy may be used to provide reliable clinical diagnosis of diseases and disorders.

The author gratefully acknowledges the support of Fisher and Paykel Healthcare (NZ) Ltd, also to David Whiting for his helpful suggestions. Thanks are also due to A/Prof. Rapoport of the New York Medical Center, for providing the data used in this preliminary work.

References

Feature Extraction of New Imagery Tasks Using Time-Frequency and Statistical Analysis for Brain-Computer Interface

Montri PHOTHISONOTHAI† and Masahiro NAKAGAWA‡

†‡Department of Electrical Engineering, Nagaoka University of Technology
1603-1 Kamitomioka, Nagaoka, Niigata 940-2188 JAPAN
Emails: montri@pelican.nagaokaut.ac.jp, masanaka@vos.nagaokaut.ac.jp

Abstract—In this paper, we investigate the possibility of new imagery tasks into simple binary commands approach to a brain-computer interface (BCI). An analysis of imaginary tasks as ‘Yes/No’ have been proposed. Since BCI technology is very helpful for the patients who are suffering from severe motor disabilities. They are require communicate with the outside world by using an electroencephalogram (EEG) signals recording on the scalp surface. Two healthy subjects (2 males), aged 25-26 years, were volunteered to participate in the experiment. During the experiment, 8-questions were used for stimuli. The feature extraction of the event-related synchronization and event-related desynchronization (ERD/ERS) change can be explained by the slope coefficient and Euclidian distance (SCED) method. The average features distribution can plot in polar plane. The experimental results of the proposed method show the average percentage values of ERD is 57% and ERS is 73% when the subjects imagine to ‘Yes’ and ‘No’, respectively.

1. Introduction

A novel communication channel between a human brain and a computer is called a brain-computer interface (BCI). This channel has been providing not only healthy peoples but also patients who are suffering from severe motor impairments or disabilities such as muscularly disabled people, people are locked in their own bodies, Amyotrophic lateral sclerosis (ALS), brainstem stroke, cerebral palsy, and numerous other diseases that they cannot use any of the traditional methods but being cognitive intact. This affect to more than two million people all over the world [1]–[2]. They are require the alternative methods for communication around their environments with the outside world. A BCI system can be realized by either training the users to control his or her brainwaves [3]. Moreover, research and development on a BCI technology are effective in the real world applications, leading to an improved quality-of-life and reduced social costs [4]. The BCI applications which are allow the subject or user to operate the brain-actuated devices such as a virtual keyboard, a cursor movement, a computer game, the internet access, a wheelchair, a speech synthesizer, an assistive appliance, and a robot control.

An electroencephalogram (EEG) provides a measure of electrical activity from neural currents within the brain. EEG pattern changes can be utilizing in an engineering point of view to establish a BCI system. The EEG signals are measuring at the scalp surface through the electrodes or noninvasive method. The basic phenomenon changes in the amplitude of the oscillatory behavior of specific cortical areas, an amplitude (or power) increase is defined as an event-related synchronization (ERS) while an amplitude (or power) decrease is defined as an event-related desynchronization (ERD) usually we so called the ERD/ERS [2].

For related works [5]–[8], the imagination of limb movements have been proposed, the motor imagery tasks such as left- and right-hand movement are hot topic until to date. However, that motor imagery tasks require the subject prior training the EEG signals to operate with the computer and also the ability of BCI system depend on the training skill.

In this paper, we would like to propose the imagination of ‘Yes’ and ‘No’ due to use for the simple binary commands approach to noninvasive BCI. Those are the novel tasks for an alternative method do not require any training. In the experiment, we set the 8-questions for the subjects answer them. An analysis of EEG signals during imagination of answers has been investigated. We proposed the time-frequency and statistical analysis in order to reveal the ERD/ERS patterns. For the feature patterns of the ERD/ERS change can be detected by using the slope coefficient and Euclidian distance (SCED) method.

The paper proceeds as follows. Section 2 the proposed method is presented. And the experimental results in Sect. 3. Finally, discussed and concluded in Sect. 4.

2. Methods

2.1. EEG signals acquisition

The electrodes were placed according to the international 10-20 system. In this paper, the electrode positions will divide to three main areas, i.e., frontal area (F7, F3, Fz, F4, F8), central area (T3, C3, Cz, C4, T4), and parietal area (T5, P3, Pz, P4, T6) on the scalp. Also, we use the ground signal at the forehead, Fpz, and the reference signal at right ear lobe, A2, depicted in Fig. 1.

In the experiment we use a multiple channel amplifier (MEG-6116; Nihon Kohden, Tokyo, Japan), analog-to-digital converter (PC-CARD-DAS16/16; Measurement
Computing, MA, USA), and the EEG electrode paste (Z-401CE; Nihon Kohden, Tokyo, Japan) is used for impedance reducing. The EEG signals were digitized at 256 samples/sec, resolution 16 bits/sample, and signals were analog bandpass filtered between 1.5 and 100 Hz. The experimental paradigm we have tested with two males (SM1 and SM2). Their age between 25-26 years old, 40-trials data per subject (5 trials per question). At the beginning of the EEG recording session, they were instructed to be relaxed with eyes-opened for 30 seconds. The distance between a monitor and position of a chair is about 90 centimeters, the one-trial epoch of time period is 12 seconds.

2.2. Spatial filtering

Firstly, the raw EEG signals were recorded from the subjects should prepare to preprocessing. For preprocessing that EEG patterns can be detected with surface Laplacian (SL) filtering than with the unprocessed raw potentials [9]. The SL filtering reduces considerably the spatial blur of the recorded EEG signals due to the head as a volume conductor. The SL filtering of a potential field can be calculated by using two dimensional second derivative operator is given by

\[ L_x = -\nabla^2 \phi_{xy} = - \left( \frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} \right) \]  

where \( \phi \) is a function of a three dimensional space, \( x, y, z, \) in case of scalp potential. Since the distances of the nearest electrodes are small in relation to the curvature of the head, a planar geometry, \( \phi_{xy} \), can be assumed as a good approximation of the local surface. In practice the SL filtering at the electrode can be numerically estimated by the weighted summation of potentials at the electrode and its nearest neighboring surround by Hjorth’s method [9]. In this paper we have modified SL filtering that it was appropriated with the electrode positions as follows

\[ \hat{x}_i = x_i - \frac{1}{M} \sum_{j \in \mathcal{N}_i} x_j \]  

where \( \hat{x}_i \) is the output signals from SL filtering, \( x_i \) is the raw EEG signals at \( i \)-th channel, \( \mathcal{N}_i \) is an index set of the neighboring channels, and \( M \) is the constant values where \( M = 2 \) if the current channels are T5, T6, F7, and F8; \( M = 3 \) if the current channels are P3, P4, Pz, T3, T4, F3, F4, and Fz; and \( M = 4 \) if the current channels are C3, C4, and Cz.

2.3. Time-frequency analysis

This paper uses the time-frequency analysis of the ERD/ERS changing over the recoded EEG signals.

2.3.1. EEG bands filtering

This step the EEG signals were filtered out to five bands. The traditional frequency bands of EEG were filtered as \( \delta (1-4 \text{ Hz}), \theta (4-8 \text{ Hz}), \alpha (8-13 \text{ Hz}), \beta_1 (13-20 \text{ Hz}), \) and \( \beta_2 (20-30 \text{ Hz}) \). An infinite impulse response (IIR) bandpass Butterworth filter with a fourth-order was applied. The EEG filtering of each bands in time-frequency plotting shown in Fig. 3.

2.3.2. Power of signal

After filtered EEG bands, the power of signal in the imagined period was computed (10 to 12 sec). We can cal-
3. Experimental results

The output feature distribution of the SCED method from the subjects imagine to ‘Yes’ and ‘No’ are shown...
Table 1: The average percentage of the ERD/ERS for imagination of ‘Yes’ and ‘No’.

<table>
<thead>
<tr>
<th>Area</th>
<th>Task</th>
<th>δ[%]</th>
<th>θ[%]</th>
<th>α[%]</th>
<th>β1[%]</th>
<th>β2[%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal</td>
<td>Yes</td>
<td>52/48</td>
<td>9/91</td>
<td>53/47</td>
<td>19/81</td>
<td>12/88</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>33/67</td>
<td>7/93</td>
<td>49/51</td>
<td>20/80</td>
<td>15/85</td>
</tr>
<tr>
<td>Central</td>
<td>Yes</td>
<td>57/43</td>
<td>7/93</td>
<td>55/45</td>
<td>12/88</td>
<td>13/87</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>27/73</td>
<td>5/95</td>
<td>42/58</td>
<td>23/77</td>
<td>12/88</td>
</tr>
<tr>
<td>Parietal</td>
<td>Yes</td>
<td>58/42</td>
<td>6/94</td>
<td>52/48</td>
<td>14/86</td>
<td>19/81</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>47/53</td>
<td>7/93</td>
<td>47/53</td>
<td>16/84</td>
<td>17/83</td>
</tr>
</tbody>
</table>

Figure 6: Average features distribution based on the SCED method in polar plane. (a) The subjects imagine to ‘Yes’. (b) The subjects imagine to ‘No’.

In Table 1, we use the objective evaluation method by taking the number of the SCED parameters in ERD (or ERS) region divide by total number of ERD and ERS region. Since the experimental results show the obvious average percentage of the ERD was 57% and ERS was 73% in δ band at central area when the subjects imagine to ‘Yes’ and ‘No’, respectively. Feature extraction of new imagery tasks using time-frequency and statistical analysis for brain-computer interface have been proposed in this paper. For future work we will classify these results to binary commands according to the subject imagines to ‘Yes’ or ‘No’. Moreover, this imaginary tasks can serve for the patient in a hospital to communicates with the physician by using the simple binary commands without training.

4. Discussions and conclusions

In Table 1, we use the objective evaluation method by taking the number of the SCED parameters in ERD (or ERS) region divide by total number of ERD and ERS region. Since the experimental results show the obvious average percentage of the ERD was 57% and ERS was 73% in δ band at central area when the subjects imagine to ‘Yes’ and ‘No’, respectively. Feature extraction of new imagery tasks using time-frequency and statistical analysis for brain-computer interface have been proposed in this paper. For future work we will classify these results to binary commands according to the subject imagines to ‘Yes’ or ‘No’. Moreover, this imaginary tasks can serve for the patient in a hospital to communicates with the physician by using the simple binary commands without training.

Acknowledgments

This research was partially supported in part by the 21st Century COE (Center of Excellence) Program “Global Renaissance by Green Energy Revolution” and the Grant-in-Aid for Scientific Research (15500070) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The authors would like to thank the subjects who took part in this research.

References


The effect of the parameter of the Inverse Function Delayed model on the success rate of the N-Queen problem

Akari Sato†, Yoshihiro Hayakawa† and Koji Nakajima†

†Laboratory for Brainware/Laboratory for Nanoelectronics and Spintronics Research Institute of Electrical Communication, Tohoku University, Katahira 2–1–1, Aoba-ku, Sendai, 980–8577 Japan
Email: ring@nakajima.riec.tohoku.ac.jp, asaka@nakajima.riec.tohoku.ac.jp, hello@nakajima.riec.tohoku.ac.jp

Abstract—Many researches have attempted to solve NP-hard or NP-complete problems by using neural networks. However, the local minima problem remain unresolved. It has been shown that the Inverse Function Delayed (ID) model, which has a negative resistance region and can destabilize local minimum states using the influence of the negative resistance, is a strong tool to avoid the local minima problem. In this paper, we show that the global and local minima states in the energy form E of N-Queen problem is separable. Then, we confirm it by simulation, and find that there are permanent oscillating states, which do not reach the global minimum, even if making only the global minimum stable. This reason is that the state transition does not proceed and that the ID network can not search for a global minimum state globally though certain neurons keep oscillating. However, we find that α in the ID model is a parameter to control a global search and that the occurrence rate of such states is lower if parameter α in the ID model is set enough large.

1. Introduction

The N-Queen problem is one of the combinatorial optimization problems. The combinatorial optimization problem is to find the combination which has minimum or maximum cost from several combinations, and can be found in actual life. The calculation amount of such a problem is as vast as the problem size is large, hence, it is classified as NP-hard or NP-complete.

Neural networks have been used to solve combinatorial optimization problems[1]. Neural networks compute in parallel, and therefore, can compute fast. In addition, it have high versatility. It has a problem, not alwayes reaching optimal answer, that is called the local minima problem.

Although, variable neural network models have been proposed, we use the Inverse function delayed (ID) model[2]. The ID model has negative resistance in its dynamics, and it has a quality of an active unit. Hence, if the equilibrium point is set in the negative resistance region by the self connection and the external input, a unit of the ID model shows an oscillation phenomenon. Using this phenomenon, the local minimum can be destabilized. Then, the ID model can have high success rate on combinatorial optimization problems[3]. However, it does not necessary mean that the state goes to the global minimum of the potential energy when it escapes from the local minimum. Hence, it is important to find a good parameter having a high reachability of the global minimum.

In this paper, we found the parameter region which can take advantage of the quality of the ID model on N-Queen problem.

2. The ID model

2.1. Basic Equations of the ID model

The ID model is described as follows

\[ \tau \frac{du_i}{dt} + u_i = \sum_j w_{ij} x_j + h_i \]  
(1)

\[ \tau \frac{dx_i}{dt} = u_i - g(x_i) \quad (\tau_s \ll \tau) \]  
(2)

where, \( u_i \) is the internal state of neuron \( i \), \( x_i \) is the output, \( h_i \) is the bias, \( w_{ij} \) is the synaptic weight from neuron \( j \) to neuron \( i \), \( \tau \) is the time constant of the internal state, \( \tau_s \) is the time constant of the output, \( g(x) \) is the inverse function of the output function, respectively.

The following formula will be obtained substituing Eqs.(1),(2) for Eq.(2) differentiated with time.

\[ \tau \frac{d^2 x_i}{dt^2} + \eta(x_i) \frac{dx_i}{dt} = -\frac{\partial U}{\partial x_i} \]  
(3)

\[ \eta(x_i) = \frac{dg(x_i)}{dx_i} + \frac{\tau_s}{\tau} \]  
(4)

\[ \frac{\partial U}{\partial x_i} = \frac{1}{\tau_s} \left( g(x_i) - \sum_j w_{ij} x_j - h_i \right) \]  
(5)

The first and the second terms on the left hand side of Eq.(3) correspond to an inertia and a friction term, and the right hand side of Eq.(3) corresponds to a force exerted by potential \( U \), respectively. The coefficient of the friction \( \eta(x) \) has a negative value when \( g(x) \) has N shape. This region that \( \eta(x) < 0 \) is called the negative resistance region.

If \( w_{ij} = w_{ji} \), we can define the energy function of the ID model the same way as in the Hopfield model, and it is as follows
\[ E = \frac{1}{2\tau} \sum_i w_{ii} x_i^2 - \frac{1}{2\tau} \sum_{i,j=0} \sum w_{ij} x_i x_j - \frac{1}{\tau} \sum_i h_i x_i \]

\[ + \frac{1}{\tau} \sum_i \int_0^{x_i} g(x_i) dx + \frac{\tau}{2} \sum_i (\frac{dx_i}{dt})^2 \]  

(6)

Its time derivative is derived by using Eqs. (1), (2), (4)

\[ \frac{dE}{dt} = - \sum_i \eta(x_i) \left( \frac{dx_i}{dt} \right)^2 \]  

(7)

where, \( \eta(x_i) \) is same \( \eta \) as in Eq.(4). Therefore, \( E \) is increased with time if \( \eta \) is negative. It is expected to get out from local minima, if local minima are in the negative resistance region.

2.2. \( g \) Function and the Negative Resistance Region

In this paper, we use following equation as the inverse function of the output function \( g(x) \), because we take up combinatorial optimization problems.

\[ g(x) = f^{-1}(x) = \alpha \left( x - \frac{1}{2} \right) \]  

(8)

\[ f^{-1}(x) = \left( \frac{1}{1 + e^{-\beta x}} \right)^{-1} = \frac{1}{\beta} \ln \left( \frac{x}{1-x} \right) \]  

(9)

Where, \( \beta \) is a gain parameter. Fig. 1 shows the shape of \( g(x) \). This function is made by adding a linear term to the inverse function of the general \( \text{tanh} \) type output function \( f(x) \) in order to make \( g(x) \) an N shape.

![Figure 1: g-function used in this paper](image)

\( g(x) \) is symmetric about \( (x, u) = (0.5, 0) \), therefore, the negative resistance region is also symmetric. Thus, we define \( L \) as the width of the negative resistance region, which is the distance from \( x = 0.5 \) to the verge of the negative resistance region

\[ L = \sqrt{1/4 - 1/\beta(a - \tau_i/\tau)} \]  

(10)

\( L \) is expressed as a function of \( \alpha, \beta, \tau_i/\tau \). However, in this paper, we control \( L \) by \( a \) mainly.

The following formula will be obtained if Eq.(8) is substituted for Eqs.(6)

\[ E = \frac{1}{2\tau} \sum_i (w_{ii} + \alpha) x_i^2 - \frac{1}{2\tau} \sum_i \sum_j w_{ij} x_i x_j \]

\[ + \frac{1}{\tau} \sum_i \int_0^{x_i} g(x_i) dx + \frac{\tau}{2} \sum_i (\frac{dx_i}{dt})^2 \]  

(11)

When the network stays at a stable equilibrium point, the fifth term of Eq.(11) is zero. Hence, this energy function can be the same form as Hopfield model’s [1] by changing the self connection \( w_{ii} \) and the bias \( h_i \) into the following \( W_i, H_i \).

\[ W_i = w_{ii} - \alpha, \quad H_i = h_i + \alpha/2 \]  

(12)

In other words, energy form can be independent of \( \alpha \).

3. The ID model for the N-Queen Problem

3.1. The N-Queen Problem

The N-Queen problem is one of a benchmark of combinatorial optimization problems. This problem is to place N queens with no collision on an \( N \times N \) chess board. One of the correct answers of the 8-Queen problem is shown in Fig. 2.

![Figure 2: A correct answer state of the 8-Queen problem](image)

In this paper, neuron \((p, i)\) corresponds to the square at row \( p \) and column \( i \) on the chess board. When a queen is placed at row \( p \), column \( i \), the output of neuron \( x_{pi} = 1 \), otherwise, \( x_{pi} = 0 \). The energy function is defined as follows

\[ E = \frac{A}{2} \sum_p \left( \sum_j x_{pj} - 1 \right)^2 + \frac{A}{2} \sum_i \left( \sum_q x_{qi} - 1 \right)^2 \]

\[ + B \sum_p \sum_i \sum_j q(p, j) x_{pj} x_{qi} + C \sum_p \sum_i x_{pi} (1 - x_{pi}) \]  

(13)

where,

\[ c_{qj}^{pi} = \begin{cases} 1 & \text{if } q \text{ and } p \text{ are on a skew} \\ 0 & \text{otherwise} \end{cases} \]  

(14)

The first and the second term represent constraint in order to place only one queen on each row and column, the third term represents constraint in order to place at most
one queen on a skew, the fourth term represents constant for the output value to be 0 or 1. When these three conditions are satisfied, the output shows correct answer, and the energy has minimum value, \( E = 0 \).

From Eq. (13), (14), synaptic weight to neuron \((q, j)\) from neuron \((p, i)\) \(w_{piqj}\) and bias \(h_{pi}\) is derived as follows

\[
\begin{align*}
\quad w_{piqj} &= -A(\delta_{pq} + \delta_{ij}) \\
& \quad -B(\delta_{pi1} + \delta_{piq}) \left(1 - \delta_{pq} \delta_{ij}\right) \\
& \quad + C \delta_{piq} \delta_{ij} \quad (15) \\
\\
\quad h_{pi} &= 2A - C/2 \quad (16)
\end{align*}
\]

In the case of the ID model, these are according to Eq. (12).

### 3.2. Dividing global minima from local minima

Here, we show that the ID model can divide global minima from local minima of the N-Queen problem, and only makes local minima unstable.

At first, we consider the location of the global minima.

From basic equations of the ID model (1), (2), the equilibrium point \(x_{pi}\) implements \(g(x) = W_{pi}x_{pi} + \theta_{pi}\), where, \(\theta_{pi} = \sum w_{piq}x_{qj} + h_{pi}\). By the correction of \(W, H\), the location of the equilibrium point is independent of \(\alpha\), hence, we consider the case of \(\alpha = 0\). At the global minimum, \(\theta\) is

\[
\begin{align*}
\quad \theta_{pi} &= 2A - C/2 \quad \text{(output is HIGH)} \\
\quad \theta_{pi} &\leq -C/2 \quad \text{(output is LOW)}
\end{align*}
\]

Here, \(x_1, x_0\) are defined the output of neuron when \(\theta_{pi} = 2A - C/2, \theta_{pi} = -C/2\), then, following equation is satisfied.

\[
\begin{align*}
\quad f^{-1}(x_1) &= (-2A + C)x_1 + 2A - C/2 \quad (17) \\
\quad f^{-1}(x_0) &= (-2A + C)x_0 - C/2 \quad (18)
\end{align*}
\]

Because \(f^{-1}(x)\) is symmetric about \((x, u) = (0.5, 0)\), the distance from \((0.5, 0)\) to \(x_1, x_0\) is equal.

It assumed that \(x_0 = 1 - x_1\), and is substituted for Eq. (18), we will get \(f^{-1}(1 - x_1) = (-2A + C)x_1 + 2A - C/2\).

\[
\begin{align*}
\quad f^{-1}(x) &= -f^{-1}(1 - x) \quad \text{because } f^{-1}(x) \text{ is symmetric, therefore, Eq. (17) is derived. Hence, the assumption is true, that is, } |x_0 - 0.5| = |x_1 - 0.5|. \quad \text{That is to say, if } x_1 \text{ is outside the negative resistance region and stable, all neurons are stable, because the output value of LOW level neurons are less or equal } x_0.
\end{align*}
\]

On the other hand, when the network has too many HIGH level neurons, the output value of these HIGH neurons are less then \(x_1\). Because such a neuron is restrained from unsatisfied conditions and \(h\) is smaller than global minima’s. In addition, when the network lacks HIGH neurons, the output value of the LOW neuron is more than \(x_0\) if the factors of the energy function \(A, B, C\) is set adequately. That is to say, the output vector of the local minimum has elements closer to \(x = 0.5\) than the global minimum. Therefore, global minima can be divided from local minima by the element that is closest to \(x = 0.5\) of the output vector, and local minima can be made unstable even if the width of the negative resistance region \(L\) is smaller than the necessary width to destabilize global minima.

### 4. Dependence of the Success Rate on the Width of the Negative Resistance Region

In previous section, we have shown that global minima are divided from local minima and that it can destabilize only local minima by the negative resistance effect. In this section, in order to confirm it, we show simulation result of the success rate as a function of the width of the negative resistance region \(L\) and we also calculate the distribution of equilibrium points for \(L = 0\).

We observe \(E\) for each 0.01\(\tau\). If differential \(\Delta E\) is satisfied \(\Delta E < \epsilon = 10^{-2}\) during 3\(\tau\), we regard the network to be in a stationary state. Then we determine it shows a global minimum or a local minimum. We use the uniform random numbers in the range of \(0.2 < x_0 < 0.8\) as the initial value of \(x_0\), and \(u_0\) is decided by \(u = g(x_0)\). Factors of the energy function are set to be \(A = 2, B = C = 1\).

Solid lines in Fig. 3 show the occurrence rate of global minimum and local minimum of the 8-Queen problem as a function of \(L\). Each plot is result from 50 samples. The sum of the ratio does not always have to be 1 because a permanent oscillating state exists.

The block graph in Fig. 3 shows the distribution of the equilibrium points as a function of the distance from \(x = 0.5\) to the element which is closest to \(x = 0.5\) when \(L = 0\). Black blocks and white blocks show global minima and local minima, respectively. The simulation result is obtained from 1000 samples.

![Equilibrium state distribution](image-url)
5. Dependence of Success Rate on $\alpha$

In the previous section, we have found even if $L$ is large enough to destabilize the local minima, the network does not always reach a global minimum. In this section, we investigate the reason for not reaching the global minimum through the parameter $\alpha$, because $L$ is controlled by $\alpha$.

Fig. 4 shows the time evolution of the output of each neurons when the network falls into a permanent oscillating state on $\tau,i/\tau = 0.01, \beta = 10, \alpha = 1.0(L \approx 0.373)$. Only particular neurons change the output, and the state transition does not proceed.

Fig. 5 shows the network state when $t = 17.4427[\tau]$. The black box shows firing neuron. After this state, neuron (a) becomes LOW and all neurons in the gray area in the figure are unstable, but neuron (a) becomes HIGH again. That is to say, this change of output of neuron (a) doesn’t influence the state transition of the network. That is considered one of the reasons that the state transition does not proceed and not reach the global minimum.

Then, we simulate that dependence of the time evolution of neuron (a) and (b) on $\alpha$ after neuron (a) becomes LOW. We fix the output of neurons expected for (a), (b). Initial states of neuron(a) and (b) are decided from nullcline of the ID model(Fig. 6). Neuron (a) is set point A because the start time is immediately after changing neuron (a) from HIGH (point C) to LOW (point A). Neuron (b) has been received inhibit inputs from three neurons until neuron (a) becomes LOW, hence, initial state of neuron (b) is set point B, where, $\theta_{b0} = H - 2A - B$.

Fig. 7–10 show simulation results. From the result, if $\alpha$ is large enough, neuron (b) changes output faster than neuron (a). That is to say, if $\alpha$ is large, the change of output of neuron (a) can influence the change of output of another neurons. To reach a global minimum, in addition, it is need that this step which is the change of output of a neuron influence another neuron’s output follow up. It is considered that if the period keeping neuron (b) HIGH is too short, this step is stopped. From the result, the period keeping neuron (b) HIGH is long as $\alpha$ is large.

As above, we conclude that that $\alpha$ is a parameter to control a chain reaction of state transition and that the ID network can search for a global minimum state of E when $\alpha$ is large enough.

6. conclusions

We showed that the ID model can divide global minima from local minima, and that it destabilizes only local minima by negative resistance effect on the N-Queen problem.

In computer simulation, we found that the network necessary reaches a global minimum even if all local minima are unstable. That is the permanent oscillation state, and the phenomenon in which only particular neurons change output occurs in this state.

However, we found that $\alpha$ in the ID model is a parameter to avoid this phenomenon, and that the appearance ratio of this phenomenon is lower if $\alpha$ is large enough.

References


A new three-level tree data structure for optimizing the traveling salesman problem by Lin-Kernighan heuristic

Hung D. Nguyen†, Ikuo Yoshihara†, Kunihito Yamamori‡ and Moritoshi Yasunaga‡

†Faculty of Engineering, University of Miyazaki
1-1 Gakuen Kibanadai Nishi, Miyazaki, Miyazaki 889-2192, Japan
‡Institute of Information Sciences and Electronics, University of Tsukuba
1-1-1 Tennoudai, Tsukuba 305-8573, Japan
Email: ndhung@med.miyazaki-u.ac.jp, yoshiha@cs.miyazaki-u.ac.jp,
yamamori@cs.miyazaki-u.ac.jp, yasunaga@is.tsukuba-u.ac.jp

Abstract– Lin-Kernighan (LK) is the most powerful local search for the Traveling Salesman Problem (TSP). The choice of data structure for tour representation plays a vital role in LK’s performance. This paper proposes a new three-level tree data structure for tour representation. Although this structure is asymptotically not better than the existing ones, it performs empirically better than the existing ones for TSPs having from a thousand to three million cities.

1. Introduction

Traveling Salesman Problem (TSP) is one of the most important and representative combinatorial optimization problems, because it is simple to state and widely applicable but difficult to solve. The TSP can be stated as follows: The salesman must visit a list of cities, all of whose locations are given. The salesman’s task is to find the cheapest tour connecting them all, visiting each city only once, and return to the city of origin. Cost here can be distance, time, or money etc. This paper deals only with symmetric TSPs, in which all the costs between any two cities are equal in both directions.

Since the TSP is NP-complete, any method that guarantees to find true optimal solutions requires running time growing exponentially with \( N \). (Throughout this paper, \( N \) is used to denote the size of problem.) Therefore, to attack large-scale problems, it is necessary to develop approximate algorithms that do not always aim at finding true optimal solutions but at finding quasi-optimal solutions in an acceptable running time. Among approximate algorithms proposed for the TSP, the Lin-Kernighan (LK) heuristic [1] is one of the most effective algorithms. Furthermore, most of the algorithms that supersede LK are meta-heuristics incorporating it in one or other ways, for example, iterated (chained) LK algorithms [2-4] and hybrid genetic algorithms [5-9]. Thus, improvements to LK are always of significant, since they improve not only the algorithm itself but also other algorithms incorporating it. To date, many improvements to LK have been proposed, e.g., two-level tree data structure [10], Mak-Morton move [11], and Helsgaun S-Opt basic move [12].

The LK heuristic is very complicated and has many choices for implementing it. Especially, the choice of data structure for tour representation plays an important role in LK’s performance. Traditionally, arrays have been used for problems with the size up to a thousand cities, two-level trees have been used for problems with from a thousand to a million cities, and binary trees (including splay trees) have been used for problems with a million or more cities. This paper proposes three-level trees as a new data structure for tour representation in LK. The new data structure performs empirically better than the conventional ones for TSPs with from a thousand to three million cities.

2. Conventional Data Structures

2.1. Basic Operations of Data Structures in LK

A data structure for tour representation in LK must support the following four basic operations:

(1) \( \text{Next}(T, a) \): This operation returns the successor of city \( a \) in tour \( T \).
(2) \( \text{Prev}(T, a) \): This operation returns the predecessor of city \( a \) in tour \( T \).
(3) \( \text{Between}(T, a, b, c) \): This operation returns true if city \( b \) lies between cities \( a \) and \( c \) in tour \( T \) (suppose that the direction for traversing the tour is well defined, e.g., by the \( \text{Next} \) operation). It returns false otherwise.
(4) \( \text{Flip}(T, a, b) \): This operation reverses the visiting order of the segment between cities \( a \) and \( b \) in tour \( T \).

Which data structure should be used for tour representation depends mainly on the size of the problems to be solved. In this section, we will briefly discuss three conventional data structures: arrays, two-level trees, splay trees, which are most commonly used for representing tours in modern LK implementations. More details about these data structures can be found in [13].

2.2. Arrays

In the array data structure, the tour is represented by two one-dimensional arrays of length \( N \) and a “reversed bit”. The first array contains the names of the cities in
visiting order. The second array contains the indices of the cities. It is an inversion of the first array. If city \(i\) is to be visited at the \(j\)th visit, then the value at the \(j\)th position of the first array is \(i\), and the value at the \(i\)th position of the second array is \(j\). Note that a Flip on a segment can be implemented by calling a Flip on the complementary segment and then reversing the whole tour. Therefore, with the use of the reversed bit, which is turned on if the tour is reversed, the Flip operation can always be implemented on the shorter segment. For arrays, the three queries \(\text{Next}\), \(\text{Prev}\), and \(\text{Between}\) can be implemented in \(O(1)\) amortized time. The Flip operation, however, takes an amortized worst-case time \(O(N)\). Therefore, this data structure is not well scalable with the size of problem. Given its small constant time overhead for the three query operations, however, arrays are suitable for solving TSPs with up to a thousand of cities.

### 2.3. Splay Trees

Another way for representing a tour is using binary search trees, with each node represents a vertex in the tour and each subtree represents a subtour. To facilitate the Flip operation, each node is attached with a reversed bit that indicates in what order the subtree rooted at that node should be traversed. If the reversed bit of a node is off (on), the subtree rooted at that node should be traversed in-order (reversed in-order). In the case that all of the four tour operations are performed by using the Splay operation [13], binary trees are called splay trees. It has been proven that all of the four tour operations of splay trees can be implemented in \(O(\log N)\) worst-case time if we always rebalance the trees after each Flip. This is the best time complexity that has been known for data structure representing tours in the LK heuristic. However, due to the large overhead, splay trees are only suitable for solving very large problems (a million or more cities).

### 2.4. Two-level Trees

The two-level tree data structure was first proposed for tour representation by Chrobak et al. [10]. The idea is to divide the tour into roughly \(N^{1/2}\) segments, each of which having a length in the range from \(N^{1/2}/2\) to \(2N^{1/2}\) cities. These segments are maintained as a doubly-linked list of nodes. It has been proven that the amortized times for performing the \(\text{Next}\), \(\text{Prev}\), and \(\text{Between}\) operations of two-level trees are all \(O(1)\), whereas the amortized worst-case time for performing the Flip operation is \(O(N^{1/2})\) if we always rebalance the tree after each Flip, or \(O(N)\) otherwise. Two-level trees are currently the fastest and most robust data structure for tour representation used in LK for solving problems ranging in size from a thousand to a million cities. Many real-world TSP applications have sizes falling in this range.

### 3. New Data Structure: Three-level TSP

We recognized that although two-level trees need to be kept roughly balance to yield the \(O(N^{1/2})\) amortized worst-case time for Flip, most of modern LK implementations do not implement rebalancing. Without rebalancing, the amortized worst-case time for doing Flip of two-level trees is \(O(N)\), the same as with arrays. The performance of two-level trees without rebalancing, however, has been proven to be much better than arrays when solving large
problems. It is probably because the amortized average-case time for doing Flip of two-level trees without rebalancing is better than that of arrays.

Based on this observation, we propose using three-level trees as a new data structure for tour representation. The three-level tree structure, as its name inspires, has 3 levels (Fig. 1). The first (children) level has $N$ elements. The second (parent) level has roughly $N^{2/3}$ elements and the third (grandparent) level has roughly $N^{1/3}$ elements. Similar to two-level trees, the elements at each level of three-level trees are maintained as a doubly-linked list. It can be easily seen that the tree operations Next, Prev, and Between of three-level trees take amortized time $O(1)$, the same as with two-level trees, but with ~1.5 times larger overheads. The amortized worst-case time for performing Flip of three-level trees is $O(N^{2/3})$ with rebalancing (since there are roughly $N^{2/3}$ elements at the second level), or $O(N)$ without rebalancing. In the case that rebalancing is not implemented, both two-level tree and three-level tree structures have the same amortized worst-case time $O(N)$ for doing Flip. However, we expect that in practice the performance of three-level trees will be better than that of two-level trees when solving large problems.

4. Experiments and Validation

4.1. Experimental Settings

Our LK implementation uses 12-quadrant nearest neighbors and quick-Boruvka initial tours for Euclidean distance problems, and 12 nearest neighbors and Nearest Neighbor initial tours for other problems. The LK search is limited to a maximum depth of 100 edges. Like many other implementations, our implementation also employs some standard improvement techniques, such as “don’t look bits” and hash tables for distance caching [14].

Our program was written in the C programming language. We only wrote the LK search engine and the three-level tree structure parts of the code. Other parts such as reading problems, building k-d trees, creating neighbor lists, generating initial tours etc. were taken from the implementation of Chained Lin-Kernighan [4]. The parts for the array, two-level tree, and splay tree structures were also taken from the above source.

TSP benchmarks, which are available at the 8th DIMACS challenge website on the TSP (http://www.research.att.com/~dsj/chtsp/), were used for our experiments. This testbed has been used for studying the asymptotic behavior of various TSP algorithms [15]. The benchmarks are divided into four classes as follows:

(1) Random uniform Euclidean class: This class includes 26 problems, with sizes increasing by a factor of $10^{0.5}$ from $10^3$ to $10^5$ cities. The $10^5$-city instance, however, was excluded as our machines did not have enough memory to handle it. Instances of this class are designed for studying the asymptotic behavior of TSP heuristics.

(2) Random clustered Euclidean class: This class includes 23 problems, with sizes increasing by a factor of $10^{0.5}$ from $10^3$ to $10^{5.5}$ cities. Instances of this class are designed to be challenging TSP heuristics.

(3) Random matrix class: This group includes 7 problems, with sizes increasing by a factor of $10^{0.5}$ from $10^3$ to $10^5$ cities. Although these instances have no direct application in practice, they offer a great challenge to many heuristics and thus are useful in studying the robustness of TSP heuristics.

(4) TSPLIB class: This class includes 11 symmetric problems from the TSPLIB [16]. The four problems, pr1002, pcb1173, rl1304, and nrw1379, are used for $N = 10^3$. The three problems, pr2392, pcb3038, and fnl4461, are used for $N = 10^{3.5}$. The two problems, pla7397 and brd14051, are used for $N = 10^4$. The two problems pla33810 and pla85900 are used for $N = 10^{4.5}$ and $N = 10^5$, respectively. These instances can be regarded as “real world” instances.

4.2. Results and Discussion

We compared the performance of LK using various data structures. Since this performance may vary from machine to machine, we used two machines for our experiments. The first was a 1.6 GHz Pentium IV running under Windows 2000 (Pentium hereafter) and the second was a Sun Ultra 80 running under Solaris 2.6 (Ultra hereafter). Both machines had 1 GB of main memory. Experimental results on the Pentium machine are given in Table I. In this table, ARY, 2LT, SPT, and 3LT denote the array, two-level tree, splay tree, and three-level tree structures, respectively. Since the preprocessing time dominates a large portion of the total running time, we measured it and the LK optimization time separately to see the differences more clearly.

It can be observed from Table I that ARY is the worst data structure for tour representation in the range being investigated. ARY wins SPT only for the $10^3$-city clustered problems. It should be noted, however, that the ARY structure suffers from a phenomenon, which has already been described in [13], that on certain types of machine, their performance for the Flip operation degrades rapidly as $N$ gets large. This phenomenon appears more severely on the Pentium than on the Ultra. For example, the Pentium is approximately two times faster than the Ultra when solving $10^3$-city uniform problems. However, it is only 1.3 times faster when $N = 10^5$ and by the time $N = 10^{3.5}$, the Ultra is even slightly faster than the Pentium. Therefore, the performance of ARY probably depends greatly on machine types.

The second observation is that 2LT wins SPT for all but the two biggest cases ($N = 10^6$ and $N = 10^{6.5}$) of uniform problems. Fredman et al. [13] reported slightly different results for their LK implementation, with 2LT still wins SPT for $N = 10^6$. This confirms the better performance of 2LT over SPT in the size range from $10^3$ to $10^6$ cities.

The third observation is the good performance of the 3LT representation. It starts to win 2LT by the time $N = 10^4$ for uniform problems, $N = 10^{4.5}$ for clustered and
TSPLIB problems, and \( N = 10^{3.5} \) for matrix problems. 3LT is \( \approx 1.6 \) times faster than 2LT when solving the 10^6.5-city problem. This supports our hypothesis that 3LT has better asymptotic performance compared to 2LT, its amortized average-case time for tour representation, seems comparable with SPT even when the number of cities grows to 10^7.

### 5. Conclusion

In this paper, three-level trees have been proposed as a new data structure for tour representation in the LK heuristic. Benchmarks ranging in sizes from 10^3 to 10^6.5 cities were used to validate the new data structure. Although the new data structure is asymptotically not superior to the conventional ones in the size range from 10^3 to 10^6.5, splay trees, which are asymptotically the best data structure for tour representation, seems unlikely to be faster than three-level trees even when the number of cities grows to 10^7.

### References

A Global Minimization Method for Separable Multimodal Functions Using a One-Dimensional Global Search

Hideo KANEMITSU1), Tao FENG2), Hideaki KONNO3), Masaaki MIYAKOSHI4) and Masaru SHIMOBO5)

1)3) Hokkaido University of Education, Hakodate 040-8567, Japan
2) Fuji Electric Information Service Co., Ltd., Yoyogi, Sibuya-ward, Tokyo 151-0053, Japan
4) Hokkaido University, Sapporo 060-8628, Japan
5) Hokkaido Information University, Ebetsu 069-8585, Japan;
E-mail: {1)hkanemt, 3)konno}@cc.hokkyodai.ac.jp, 2)feng-tao@fujielectric.co.jp
4)miyakosi@main.eng.hokudai.ac.jp, 5)shimbo@do-johodai.ac.jp

Abstract—We propose a method for solving global minimization problems with separable multimodal objective functions of \(n\)-real variables: \(\sum_{i=1}^{n} f_i(x_i)\) (hereafter called sum form) and \(\prod_{i=1}^{n} f_i(x_i)\) (hereafter called product form) and bounded constraints. In order to find a global minimum in a few function evaluations, a one-dimensional global search is repeatedly used in this method. We show that the global minimum can always be found by a one-dimensional global search in separable objective functions with respect to sum and product forms, and we propose an algorithm for searching for the global minimum of these functions. We show by results of numerical experiments using simple test functions that the algorithm effectively finds the global minima in only a few function evaluations.

1. INTRODUCTION

In recent years, many deterministic [2] and stochastic algorithms [7] have been proposed for solving a global minimization problem (P) of a multivariate function with bounded constraints:

\[
\begin{align*}
\text{(P)} & \quad \text{Minimize } f(x) \equiv f(x_1, x_2, \ldots, x_n), \\
& \quad \text{Subject to } x_i \in [L_i, U_i], \quad i = 1, 2, \ldots, n, \\
& \quad D^n = \{ x \in R^n | x_i \in [L_i, U_i], \quad i = 1, 2, \ldots, n \}.
\end{align*}
\]

Deterministic algorithms [2] repeatedly divide a given region into subregions, select a subregion in which a global minimum is expected to be included, and give a guarantee of successfully finding the global minimum under highly restrictive conditions on objective functions (for example, Lipschitz continuity with a known Lipschitz constant). However, in this method, the algorithms become complicated with exponential increase in the number of variables.

Deterministic and stochastic methods have the common problem of exponential increase in computational complexity with increase in the number of variables. Such a searching problem is known as "curse of dimensionality," and the problem has not been able to be overcome for a long time.

In this paper, we focus on global minimization problems with two types of separable objective functions. The first type of function (sum form) is called an ordinary separable function, and it enables a global minimum to be found by repeating a one-dimensional global search [6]. Sohdoji [6] reported that if there exists a term such that \(f_j(x_j) < 0\) in the product form \(f(x) = \prod_{i=1}^{n} f_i(x_i)\), this problem cannot be solved easily as it can be in the case of the sum form. Therefore, we mainly consider a property of and an algorithm for the second type of problem (product form) after showing a property of and an algorithm for the first type of problem.

The remainder of the paper is organized as follows. Definitions of the two types of separable problems, notations and a one-dimensional global search algorithm are given in section 2. In section 3, properties of and algorithms for both types of problems are described. Results of tests using the proposed algorithms are shown in section 4. Finally, concluding remarks are presented.

2. PRELIMINARIES

2.1. Problems and Assumptions

In this paper, we describe two types of global minimization problems, \(P_S^k\) and \(P_S^k\), that have separable objective functions with respect to the sum and the product forms:

\[
\begin{align*}
\text{(P}_S^k\text{)} & \quad \text{Minimize } f(x) \equiv f(x_1, x_2, \ldots, x_n) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) = \sum_{i=1}^{n} f_i(x_i), \\
& \quad \text{Subject to } x_i \in [L_i, U_i], \quad i = 1, 2, \ldots, n, \\
& \quad D^n = \{ x \in R^n | x_i \in [L_i, U_i], \quad i = 1, 2, \ldots, n \}.
\end{align*}
\]

\[
\begin{align*}
\text{(P}_S^k\text{)} & \quad \text{Minimize } f(x) \equiv f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n) = \prod_{i=1}^{n} f_i(x_i), \\
& \quad \text{Subject to } x_i \in [L_i, U_i], \quad i = 1, 2, \ldots, n, \\
& \quad D^n = \{ x \in R^n | x_i \in [L_i, U_i], \quad i = 1, 2, \ldots, n \}.
\end{align*}
\]

Assumption 1 In both problems \(P_S^k\) and \(P_S^k\), suppose each term \(f_i(x_i)\) of the objective function \(f(x)\) is continuous.
Assumption 2 Each problem has a finite number of isolated local minima \( x^{*j} \in \mathbb{D}^j \) (\( j = 1, 2, \ldots, M \)) and of isolated local maxima \( x^{*k} \in \mathbb{D}^k \) (\( k = 1, 2, \ldots, M \)), and both sets of local minima and maxima and both sets of function values at local minima and maxima are written by
\[
\begin{align*}
\mathcal{X}_x &= \{ x_1^{*}, x_2^{*}, \ldots, x_M^{*} \}, \\
\mathcal{F}_x &= \{ f(x_1^{*}), f(x_2^{*}), \ldots, f(x_M^{*}) \}.
\end{align*}
\]

Moreover, there is a unique local minimum \( \tilde{x}^{*} \) and its function value \( \tilde{f}^{*} \) and there is a unique local maximum \( \tilde{x}^{**} \) and its function value \( \tilde{f}^{**} \) in each problem as follows:
\[
\begin{align*}
\tilde{x}^{*} &= \min \{ x_i^{*} \}, \\
\tilde{f}^{*} &= \min \{ f(x_i^{*}) \}, \\
\tilde{x}^{**} &= \max \{ x_i^{*} \}, \\
\tilde{f}^{**} &= \max \{ f(x_i^{*}) \}.
\end{align*}
\]

2.2. Notation

In order to separate input and output of an algorithm, we adopt the following notation.

Notation 1 An algorithm is expressed by
\[
(r_1, r_2, \ldots, r_q) \leftarrow \text{name} (a_1, a_2, \ldots, a_p).
\]

This notation means that the algorithm \text{name} with input \( p \)-arguments \( (a_1, a_2, \ldots, a_p) \) is applied and then results \( (r_1, r_2, \ldots, r_q) \) are obtained.

2.3. One-dimensional Global Search

Since a one-dimensional global search method [3] was commonly used in our proposed methods, we describe the algorithm as a preliminary.

The algorithm finds the global minimum \( \tilde{x}^{*} \) and its function value \( \tilde{f}^{*} \) for an objective function \( f(x) \) on a closed interval \([L, U]\) for a given step size \( h \). The steps of the algorithm are as follows.

\[
(\tilde{f}^{*}, \tilde{x}^{*}) \leftarrow \text{G-OneDimSearch} (f, L, U, h);
\]

GL1. [Initialize]
\[
\begin{align*}
X_b &\leftarrow \emptyset; \quad F_b &\leftarrow \emptyset; \quad N \leftarrow \lfloor (U - L) / h \rfloor; \\
f^{(0)} &\leftarrow f(L); \quad f^{(1)} &\leftarrow f(L + h); \quad \tilde{f}^{**} \leftarrow \infty.
\end{align*}
\]

GL2. [Find three neighbor points with bracketing of a local minimum]
\[
\text{for } j = 2 \text{ to } N \text{ do}
\begin{align*}
x^{(j)} &\leftarrow L + j \cdot h; \quad f^{(j)} = f(x^{(j)}); \\
\text{if } f^{(j-2)} &> f^{(j-1)} \text{ and } f^{(j-1)} < f^{(j)} \text{ then} \\
X_b &\leftarrow X_b \cup \{ x^{(j-2)}, x^{(j-1)}, x^{(j)} \}; \\
F_b &\leftarrow F_b \cup \{ f^{(j-2)}, f^{(j-1)}, f^{(j)} \}; \\
\text{if } f^{(j-1)} &< \tilde{f}^{**} \text{ then} \tilde{f}^{**} &\leftarrow f^{(j-1)};
\end{align*}
\]

od;

GL3. [Apply local minimization]

Apply local minimization to three neighbor points and their function values in the set \( X_b \) and the set \( F_b \) except for three neighbor points and their function values for which lower bound estimations of the function values are greater than \( \tilde{f}^{**} \). Let the lowest function value in the obtained local minima be \( \tilde{f}^{**} \) and its point be \( \tilde{x}^{**} \).

Since the algorithm \( G_{\text{OneDimSearch}}(\cdot) \) that finds the global maximum \( \tilde{x}^{**} \) and its function value \( \tilde{f}^{**} \) is similar to \( G_{\text{OneDimSearch}}(\cdot) \), we omit steps of the algorithm.

In a one-dimensional global search, the following property holds[3].

Property 1 Let the lower unimodal region of the global minimum \( x^{**} \) of \( f(x) \) on an interval \([L, U]\) be \( Ru(x^{**}) \), where \( Ru(x^{**}) \) is defined as the maximum region \([a, b]\) such that the following condition is satisfied:
\[
\begin{align*}
f(x) &> f(y), \quad a \leq y < x^{**} \\
f(x) &< f(y), \quad x^{**} < y \leq b.
\end{align*}
\]

Then, if \( h \leq 1/2 \cdot \min \{ x^{**} - a, b - x^{**} \} \) holds, the algorithm \( G_{\text{OneDimSearch}}(\cdot) \) always finds the global minimum of function \( f(x) \).

An example of unimodal region \( Ru(x^{**}) \) and step size \( h \) is shown in Figure 1.

![Figure 1: Unimodal region Ru(x***) and step size h](image-url)

3. PROPERTIES OF BOTH PROBLEMS AND THESE ALGORITHMS

In this section we show properties and algorithms of both separable problems \( P^*_S \) and \( P^*_S \).

3.1. Separable Problem \( P^*_S \) w.r.t. the Sum

Let us describe the problem of minimizing a separable objective function \( f(x) \) with respect to the sum in a global
minimization problem $P_5^*$. We have

$$
\min\{f(x)\} = \min\{f_1(x_1)\} + \min\{f_2(x_2)\} + \cdots + \min\{f_n(x_n)\}
= \sum_{i=1}^{n} \min\{f_i(x_i)\}.
$$

From equation (1), we can show the following property.

**Property 2** If an objective function has a separable form with the sum, the global minimum can be found by repeatedly minimizing each term $f_i(x_i)$ ($i = 1, 2, \ldots, n$).

From the above properties, an algorithm using a one-dimensional global search $G_{OneDimSearch}$ that finds the global minimum $x^{**}$ and its function value $f^{**}$ for an objective function $f$ on a searching region $D^* = \prod_{i=1}^{n} [L_i, U_i]$ and for a given step size $h = (h_1, h_2, \ldots, h_n)$ can easily construct as follows.

$$(f^{**}, x^{**}) \leftarrow G_{min}\{f(x)\}. \quad \text{(f, D^n, h)};$$

**GSS1. [Initialize]**

for $i = 1$ to $n$

$$
(f_i^{**}, x_i^{**}) \leftarrow G_{OneDimSearch}(f_i, L_i, U_i, h_i);
$$

**GSS2. [Set global minimum $x^{**}$ and its function value $f^{**}$]**

$$
x^{**} \leftarrow (x_1^{**}, x_2^{**}, \ldots, x_n^{**}); \quad f^{**} \leftarrow f_1^{**} + f_2^{**} + \cdots + f_n^{**};
$$

From Property 1 and Property 2, if $h_i \leq 1/2 \cdot \min\{(x_i^{**} - a_i, b_i - x_i^{**})\}$ ($i = 1, 2, \ldots, n$) holds, the algorithm can always find the global minimum of the function $f(x)$ with constraints $L_i \leq x_i \leq U_i$ ($i = 1, 2, \ldots, n$).

Moreover, this algorithm can find the global minimum by using $n$-times of one dimensional global searches.

### 3.2. Separable Problem $P_5^*$ w.r.t. the Product

In order to investigate a separable objective function $f(x)$ with respect to the product in a global minimization problem $P_5^*$, we discuss the problems by separating two cases.

If all minima of $f_i$ ($i = 1, 2, \ldots, n$) are greater than or equal to zero (that is, $f^{**} = \min_{i=1,\ldots, n}\{f_i(x_i)\} \geq 0$), then the following equation, which is similar to equation (1), holds.

$$
f^{**} = \min\{f(x)\} = \min\{f_1(x_1)\} \cdot \min\{f_2(x_2)\} \cdots \min\{f_n(x_n)\}
= \prod_{i=1}^{n} \min\{f_i(x_i)\}.
$$

As another case, if some minima of $f_i$ has a negative value (that is, $f^{**} = \min_{i=1,\ldots, n}\{f_i(x_i)\} < 0$), then the following equation holds.

$$
f^{**} = \min\{f(x)\} = \min\{\min\{f_1(x_1)\}, \max\{f_1(x_1)\}\}
\cdot \min\{f_2(x_2)\} \cdot \max\{f_2(x_2)\}
\cdots \min\{f_n(x_n)\} \cdot \max\{f_n(x_n)\}
= \min\left\{ \prod_{i=1}^{n} \min\{f_i(x_i)\}, \prod_{i=1}^{n} \max\{f_i(x_i)\} \right\}.
$$

The above equation denotes the minimum in all products of $2^n$-combinations such that it selects one element from the set $\{\min\{f_i(x_i)\}, \max\{f_i(x_i)\}\}$.

From the above property, an algorithm that finds the global minimum $x^{**}$ and its function value $f^{**}$ for an objective function $f$ on a searching region $D^* = \{x \in \mathbb{R}^n | x_i \in [L_i, U_i]\}$ and for a given step size $h = (h_1, h_2, \ldots, h_n)$ can be constructed as follows.

$$(f^{**}, x^{**}) \leftarrow G_{min}\{f(x)\}, \quad (f^{**}, h);$$

**GSP1. [Initialize]**

for $i = 1$ to $n$

$$
(f_i^{**}, x_i^{**}) \leftarrow G_{OneDimSearch}(f_i, L_i, U_i, h_i);
$$

**GSP2. [Set global minimum $x^{**}$ and its function value $f^{**}$]**

$$
(f^{**}, x^{**}) \leftarrow G_{min}\{f(x)\}, \quad (f^{**}, h);
$$

for $i = 1$ to $n$

$$
(f_i^{**}, x_i^{**}) \leftarrow G_{OneDimSearch}(f_i, L_i, U_i, h_i);
$$

Here, this algorithm can find the global minimum by using $n$-times of one dimensional global searches.

### 4. SIMPLE NUMERICAL EXPERIMENTS

Our two proposed algorithms were implemented in programming language C (gcc 3-3-3), and all numerical experiments for which results are shown in this paper were carried on an IBM note PC Think Pad X21 (600MHz Pentium III) with double precision.

(1) Rastrign’s function of $n$-variables (Separable problem $P_5^*$ w.r.t. the sum)

$$
f(x) = 10n + \sum_{i=1}^{n} (x_i^2 - 10\cos(2\pi x_i)),
$$

$$
x_i \in [-5.12, 5.12], \quad (i = 1, 2, \ldots, n),
x^{**} = (0, 0, \ldots, 0), \quad f^{**} = 0
$$

877
Table 1: Comparison of former methods and our method with respect to the number of function evaluations $N_{fe}$ of $n$-variables in Rastrigin function

<table>
<thead>
<tr>
<th>$n$</th>
<th>authors (year)</th>
<th>method</th>
<th>$N_{fe}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Our method</td>
<td>Opt_seph</td>
<td>490</td>
</tr>
<tr>
<td>15</td>
<td>Our method</td>
<td>Opt_seph</td>
<td>735</td>
</tr>
<tr>
<td>20</td>
<td>Ono, I. et al. (2000)[5]</td>
<td>UNGX-GA</td>
<td>$\approx$ 3,400,000</td>
</tr>
<tr>
<td>20</td>
<td>Our method</td>
<td>Opt_seph</td>
<td>980</td>
</tr>
</tbody>
</table>

Table 2: Comparison of former methods and our method with respect to the number of function evaluations $N_{fe}$ of $n$-variables in Tsuda function

<table>
<thead>
<tr>
<th>$n$</th>
<th>authors (year)</th>
<th>method</th>
<th>$N_{fe}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>Tsuda, T. (1977)[8]</td>
<td>Interactive method</td>
<td>72,962</td>
</tr>
<tr>
<td>5</td>
<td>Our method</td>
<td>Opt_seph</td>
<td>150</td>
</tr>
</tbody>
</table>

In this experiment, we set a parameter of step size $h_i$ of $G_{OneDimSearch}()$ as $h_i = 0.25$ ($i = 1, 2, \ldots, n$).

Numerical results for $n = 10, 15, 20$ and comparison of our method and other methods are shown in Table 1. Our method found the global minimum by several hundred to several thousand fewer function evaluations $N_{fe}$ than those required by the other methods.

(2). Tsuda’s function of 5-variables [8](Separable problem $P_S$ w.r.t. the product)

$$f(x) = f_1(x_1)f_2(x_2)f_3(x_3)f_4(x_4)f_5(x_5),$$

where

$$f_1(x_1) = x_1(x_1 + 13)(x_1 - 15)/100,$$
$$f_2(x_2) = (x_2 + 15)(x_2 + 1)(x_2 - 8)/100,$$
$$f_3(x_3) = (x_3 + 9)(x_3 - 2)(x_3 - 9)/100,$$
$$f_4(x_4) = (x_4 + 11)(x_4 + 5)(x_4 - 9)/100,$$
$$f_5(x_5) = (x_5 + 9)(x_5 - 9)(x_5 - 10)/100.$$  

$x_i \in [-10, 10], (i = 1, 2, \ldots, 5),$

$x^{**} = (8.7564, -9.3582, -4.5720, 3.5921, -2.8400),$

$f^{**} = -24416.03$

In this experiment, we set a parameter of step size $h_i$ of $G_{OneDimSearch}()$ as $h_i = 2.0$ ($i = 1, 2, \ldots, n$).

Numerical results and comparison of our method and other methods are shown in Table 2. Our method found the global minimum by several hundred fewer function evaluations $N_{fe}$ than those required by other methods.

5. CONCLUDING REMARKS

We have proposed two methods for solving global minimization problems with separable multimodal objective functions of $n$-real variables and bounded constraints. We have shown that the global minimum can always be found by $n$ or $2n$ times of one-dimensional global searches in separable objective functions with respect to the sum and product forms, and we have also proposed two algorithms for searching for the global minimum of these functions. We have shown by results of numerical experiments in simple test functions that the algorithm effectively finds the global minimum in only a few function evaluations. Moreover, by combining these two algorithms, mixed separable problems such as $f(x) = \sum_{i=1}^{n} f_i(x_i) + \prod_{j=1}^{n} g_j(x_j)$, $\sum_{i=1}^{n} (\prod_{j=1}^{i} f_j(x_j))$ and $\prod_{i=1}^{n} (\sum_{j=1}^{i} f_j(x_j))$ can be solved.

References


Cooperative Behavior in Coupled Simulated Annealing Processes with Variance Control

Samuel Xavier-de-Souza, Johan A.K. Suykens, Joos Vandewalle, and Désiré Bollé

Abstract—In this paper we describe the use of coupling to interconnect different Simulated Annealing (SA) processes. The objective is to allow cooperative behavior among the processes in order to improve performance. Coupled Simulated Annealing (CSA) permits a high degree of parallelization while delivering much better results than a typical Parallel SA (PSA) algorithm. This is possible due to the introduction of coupling in the acceptance probability functions. Moreover, the coupling also allows controlling the variance of the acceptance probabilities. This is especially important because it reduces the sensitivity to initial parameters, while guiding optimization to quasi-optimal runs. It was observed that the solutions generated by CSA are more concentrated around the global optimum, while PSA has concentrations of solutions often in unfavorable regions of the cost function. Also, the number of iterations per process necessary to reach a given minimum energy tolerance decreases exponentially when the number of optimizers is increased.

1. Introduction

Coupling has been applied in recent years to many different engineering applications and methods. It has proven to be helpful for synchronizing chaotic dynamical systems [6], where two identical systems can synchronize by coupling one of their state variables. Such approaches have attracted the interest of many researchers which found applications in many different fields, including communication and robotics.

Coupling has also been applied to develop the concept of Cellular Neural Networks (CNNs) [2], which features many dynamical systems, or cells, disposed in a regular grid with coupling between neighbor cells. Nowadays, CNNs gave the researchers in the field, another paradigm for information processing which can take advantages of massively parallel VLSI implementations to process data at extraordinary speed. This is only possible due to the clever yet simple coupling applied to cells.

In optimization, coupling is used to aid gradient based methods to escape from local minima [8]. In that approach, coupling of local optimization processes outperforms multi-start methods by minimizing the average cost of all coupled processes, subject to synchronization constraints between the solutions of the individual processes.

In this paper, we analyze the performance and behavior of an algorithm called Coupled Simulated Annealing (CSA). Basically, it features several SA processes running in parallel, coupled by their acceptance probabilities. The coupling provides CSA with the ability to perform optimization which has a good balance between localized search and global exploration for multi-modal problems having several local minima. Moreover, the structure of CSA permits the variance control of the acceptance probabilities via the acceptance temperature. This not only improves the performance, but also reduces the sensitivity of the algorithm to the initial acceptance temperature.

This paper is organized as follows. In Section 2, we describe the CSA algorithm and its characteristics, and we present the variance control of the acceptance probabilities. The results of our experiments are presented in Section 3, followed by the conclusion.

2. Coupled Simulated Annealing

While in classical Simulated Annealing [4] the acceptance probability of an uphill move is often given by the Metropolis rule, which depends only on the current and the probing solution, in CSA the decision of accepting such a move takes into account other current solutions. Namely, this probability depends also on the costs of the solutions in a set $\Theta \in \Omega$, where $\Omega$ is the set of all possible solutions. This dependence is given by the coupling term $\gamma$ which is generally a function of the costs of the solutions in $\Theta$. In CSA, the acceptance probability $A_\Theta$ and the coupling term $\gamma$ are given by

$$A_\Theta(y, x_i \rightarrow y_i) = \exp\left(\frac{E(x_i) - \max_{z \in \Theta} E(z)}{T^c_k}\right) \gamma,$$  \hspace{1cm} (1)

with

$$\gamma = \sum_{y \in \Theta} \exp\left(\frac{E(x) - \max_{z \in \Theta} E(z)}{T^c_k}\right).$$  \hspace{1cm} (2)
Here $T^a_k$ is the acceptance temperature and $x_i$ and $y_i$ denote an individual solution of $\Theta$ and its corresponding probing solution, respectively. These two equations define $A_\theta$ as a probability. The sum of probabilities of leaving any of the current states equals 1. Fig. 1 depicts the main differences between SA and CSA.

Functionally, CSA differs from an ensemble of SA processes [5, 7] because of two aspects. The first is the coupling which modifies the acceptance probability of each process according to the energy of the current solutions of all processes. The other aspect is blind acceptance. While downhill moves are always accepted during the optimization process, in CSA, the decision to accept an uphill move does not depend on the destination of the move, or target solution. In other words, $A_\theta(y, x \rightarrow y)$ is not a function of $y_i$. At first sight, this property may not seem helpful because the target solution may be much worse than the original one. Moreover, it may be argued that an excellent solution can easily be lost with such an approach. However, there are good reasons to use blind acceptance in CSA. The first one is that blind acceptance of uphill moves improves exploration of the energy surface. This is an essential property to solve hard multi-modal optimization problems. Additionally, due to the coupling of the acceptance functions of the different processes, uphill moves are much more likely to happen in processes that hold poor solutions than with processes holding good solutions. Therefore, the best solutions are not easily lost, whereas the poorest are fastly discarded.

The acceptance temperature in CSA is not responsible for weighting the difference between the energy of the probe and current solutions as it is in classical SA, but rather it is responsible for weighting the proportion that each acceptance probability has to the overall sum of the acceptance probabilities in $\Theta$, which in any case must be equal to 1 as guaranteed by (1) and (2). This temperature can then be used to control the variance of the probabilities regardless of the current energies. Although the ideal variance value is unknown to us, our experiments with different cost functions show that values in the neighborhood of the maximum variance deliver the best results. Typically, we recommend $99\%$ of the maximum variance value. A very simple control rule can be used to steer this variance to the desired value. It can be done in the following manner:

$$
\begin{align*}
\text{if } \sigma^2 < \sigma^2_D, \text{ then, } & T^a_k = T^a_{k-1} (1 - \alpha), \\
\text{if } \sigma^2 > \sigma^2_D, \text{ then, } & T^a_k = T^a_{k-1} (1 + \alpha),
\end{align*}
$$

where $\sigma^2_D$ is the desired variance value and $\alpha$ is the rate for the increase or decrease of the temperature, typically in the range of $(0, 0.1]$. If the value of the acceptance variance is below its desired value, the acceptance temperature is decreased by a factor of $1 - \alpha$, otherwise, it is increased by a factor of $1 + \alpha$. Such simple variance control can be applied only due to the coupling in the acceptance probability function. It substitutes a schedule for the acceptance temperature and more importantly, it works for any initial acceptance temperature. This is important because the setup of initial parameters in SA is most of the time a very cautious work. With this approach, we eliminate two initialization aspects at once, which are the choices for an acceptance schedule and an initial acceptance temperature. In return, two other parameters are introduced, $\alpha$ and $\sigma^2_D$, but these have a well defined operating range and are much less dependent on the optimization problem at hand.

3. Experiments and Results

We have tested CSA and the variance control explained above in a set of multi-modal functions with dense local minima. For comparison, we have also performed experiments using CSA without Variance Control (CSAwoVC) and using a Parallel SA (PSA) [1] algorithm. This parallel version of SA features several sequential SA processes running in parallel and sharing the best current solution. As soon as one of the parallel instances finds a better solution, all the others are informed about the new current solution and proceed to the next generation step.

3.1. Test Problems

We have used a set of four $D$-dimensional functions as test problems for the algorithms under analysis. All four functions share the property of being multi-modal with dense and uniform sets of local minima and are described by the following equations.

$$
\begin{align*}
f_0(x) &= 1 - \prod_{i=1}^D \text{sign}(x_i) \left(\frac{\sin x_i}{x_i}\right)^{\frac{1}{2}}, \\
f_2(x) &= \sum_{i=1}^D \left[ x_i^2 - 10\cos(2\pi x_i) + 10 \right], \\
f_3(x) &= -20 \exp\left(-0.2 \sqrt{\frac{1}{D} \sum_{i=1}^D x_i^2} \right) \\
&\quad - \exp\left(\frac{1}{D} \sum_{i=1}^D \cos(2\pi x_i) \right) + 20 + e, \\
f_4(x) &= \frac{1}{1000} \sum_{i=1}^D x_i^2 - \prod_{i=1}^D \cos\left(\frac{x_i}{\sqrt{\theta}}\right) + 1.
\end{align*}
$$
3.2. Results and Discussion

We have performed a variety of experiments in order to assess the performance of CSA and analyze its behavior. For all performed experiments, we have used the schedule in [9] for the generation procedure of all tested algorithms. For the acceptance procedure, we used the schedule in [3] for the PSA and CSAwoVC algorithms. For CSA, the temperature was used to perform the control of the acceptance probabilities.

The coupling in CSA has the objective to increase cooperation among optimizers and provide better acceptance decisions. This ensures that a process in a higher energy region concentrates on exploration rather than on localized search, as opposed to a process in a lower energy region. Such effect does not exist in PSA, where the search is concentrated in isolated spots of the energy surface. This concentration only changes its focus when a better solution is found. Fig. 3 has illustrative samples of typical PSA and CSA runs. Every sub-plot in the figure depicts the spread of the solutions visited by each algorithm. The global optimum lies exactly in the center of each plot. It can be seen that while PSA features many concentrated sets of solutions sparsely distributed, which are not necessarily around the global optimum, CSA features mesh-like concentrations which are around the center of the plot. It can also be seen that although each plot has the same number of visited solutions, CSA plots seems to fill better the solution space. We have observed the same results for other test problems. These results suggest that CSA has very good exploration characteristics, especially when qualitatively compared with PSA.

Fig. 4 shows the performance of CSAwoVC versus PSA for our four test problems. Both algorithms had their initial temperature obtained by an exhaustive search. It can be seen that CSAwoVC has a much better performance, except for test function $f_4$.

We tested the effect of the variance control in the performance of CSA by running an experiment that compares it with CSAwoVC, and two PSA setups for test function $f_2$, with $D = 5$ and $m = 5$. The experiment consists of 1000 runs of each algorithm with 1000 iterations, which is not much but it is enough to show how the algorithms behave in the early stages of the optimization. The acceptance temperature varied from 0.01 at iteration 1 until twice the mean energy at iteration 1000, except in one of the PSA setups, which had fixed $T_0 = 0.1$. Box-plots of the results are
shown in Fig. 5. The reader can observe that for the PSA setup in which we used the variation in the initial acceptance temperature, the performance was much poorer than the one of the other algorithms. This happens because the performance of PSA, as of many other SA algorithms, depends substantially on the initialization parameters. For the PSA with fixed initial acceptance temperature, the results improved much. Nevertheless, both CSA algorithms performed better. Besides the superior performance of CSA with the variance control w.r.t. all other algorithms, it presented also the smallest variance in the results. A zoomed version of the box-plots for the first 3 algorithms can be seen in the inner plot of Fig. 5.

At last, we performed experiments with CSA to check the scaling of the necessary number of iterations to reach a given minimum energy tolerance, with an increase in the number of optimizers. These tests were executed for function $f_2$ with several different values for $D$. The results can be seen in Fig. 6, which is presented with a logarithmic scale in the vertical axis for better visualization. This figure suggests that an increase in the number of optimizers decreases exponentially the number of necessary iterations to reach a given energy tolerance, regardless of the dimension $D$ of the problem.

4. Conclusion

We have described the algorithm Coupled Simulated Annealing (CSA). In this algorithm, several Simulated Annealing (SA) processes are coupled by their acceptance probabilities. Additionally, we have presented a straightforward control rule for the variance of the acceptance probabilities among the different processes. The results confirmed the positive effect of the cooperation introduced by the coupling in the search capabilities of the CSA algorithm.

Acknowledgments Research supported by: • Research Council KUL: GOA-Mefisto 666, GOA-AMBSoRICS, BOF OT/03/12, Center of Excellence Optimization in Engineering; • Flemish Government: o FWO: PhD/postdoc grants, G.0407.02, G.0080.01, G.0211.05, G.0499.04, G.0226.06, research communities (ICCoS, ANMMM); o Tournesol 2005 • Belgian Federal Science Policy Office IUAP P5/22. J. Suykens is an associate professor, D. Bollé and J. Vandewalle are full professors, all with K.U. Leuven Belgium. DB thanks Marcus Müller for useful correspondence, and SXS thanks João Ramos and Muştak E. Yalcın for insightful discussions.

References

Metaheuristic method by using a chaos generator with sinusoidal perturbations for global optimization

Keiji Tatsumi†, Yoshinori Obita† and Tetsuzo Tanino†

†Graduate School of Engineering, Osaka University
2-1 Yamada-oka, Suita, Osaka 565-0871 JAPAN
Email: tatsumi@eei.eng.osaka-u.ac.jp, tanino@eei.eng.osaka-u.ac.jp

Abstract—Recently, global optimization methods using chaotic dynamics have been investigated. In those methods, it is significant what kind of chaotic dynamical system is used. However, the system used in the existing method for generating a chaotic sequence has some drawbacks. In this paper, in order to improve them, we propose a new dynamical system which generates a chaotic sequence by the steepest descent method for minimizing an objective function with additional sinusoidal perturbation terms. We compare the proposed method with the existing method through computational experiments by applying them to some global optimization problems.

1. Introduction

Mathematically, the chaos means an aperiodic deterministic behavior which is exceedingly sensitive to its initial conditions. Even though the model of the system is well defined and contains no random parameters, the behavior appears to be random. In the field of optimization, such behavior is exploited in some metaheuristic methods [1, 2, 4] for the global optimization problem, which has a large number of local minima. Those methods avoid being trapped in an undesirable local minimum by making use of the chaotic behavior, and aim to find a desirable solution within a practical time. For them, it is extremely important what kind of chaos is used. A chaotic sequence generated by the existing method often tends to overconcentrate around the boundary of the feasible region. In addition, there exist some intervals of a bifurcation parameter called windows, in which any sequence converges to several periodic points. Hence, we have to select appropriate parameters to avoid the windows.

Therefore, in this paper, we propose a new chaos generator to overcome the drawbacks. It which is the steepest descent model for the objective function with additional sinusoidal perturbation terms. Moreover, we show sufficient conditions for parameters under which the proposed model generates a chaos sequence.

2. Marotto’s chaos

In this paper, we focus on the chaos in the sense of Marotto which appers in a dynamical system having some kind of fixed point called the snap-back repeller. We introduce it in this section.

Let us consider a discrete-time system as:

\[ (S) \ x(t + 1) = f(x(t)), \]

where \( x(t) \in \mathbb{R}^n, t = 1, 2, \ldots \) and \( f : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is continuously differentiable. Let \( f^{(p)} \) denote the composition of \( f \) with itself \( p(>0) \) times, then a point \( x \) is called a \( p \)-periodic point if \( f^{(p)}(x) = x \) but \( f^{(k)}(x) \neq x \) for all \( k \) such that \( 1 \leq k < p \). In particular, a point \( x \) satisfies \( f(x) = x \) is called a fixed point. The \( \varepsilon \)-neighborhood \( N_\varepsilon(x) \) of a point \( x \) denotes the set of points inside a ball with center \( x \) and radius \( \varepsilon > 0 \), i.e., \( N_\varepsilon(x) := \{ y \in \mathbb{R}^n \mid ||x - y|| < \varepsilon \} \). Then, we introduce the following definition:

**Definition 1** A fixed point \( x^* \) of system \( S \) is called the snap-back repeller of \( S \), if there exists a positive real number \( \varepsilon \) and \( x^0 \in N_\varepsilon(x^*) \) with \( x^0 \neq x^* \) such that all eigenvalues of \( \nabla f(x) \) exceed the unity in norm for all \( x \in N_\varepsilon(x^*) \) and \( f^{(m)}(x^0) = x^* \) with \( \det(\nabla f^{(m)}(x)) \neq 0 \) for some positive integer \( m \).

Then, it is well known that the existence of a snap-back repeller in the system implies the chaos, as follows.

**Theorem 1** Marotto Theorem [5]

If system \( S \) has a snap-back repeller then the system \( S \) is chaotic. In other words, there exist:

1. a positive integer \( N \) such that \( S \) has a \( p \)-periodic point for any \( p \geq N \).
2. a scrambled set \( S \subset \mathbb{R}^n \) of \( f \).

In this paper, we call a scrambled set generated by the chaotic system also chaotic.

3. Chaos generator with transformation of variables

In this paper, we deal with the following nonlinear optimization problem,

\[ (P1) \min f(x) \quad \text{s.t.} \quad x \in X, \]

where \( X \subset \mathbb{R}^n \) is a closed set, and \( f \) is a multi-modal continuously differentiable function. Since this problem
has a large number of local minima, it is difficult to find the global optimal solution. Thus, in order to solve this problem, metaheuristic methods making use of the chaotic dynamics have been investigated [1, 2, 4]. In those methods, the representative technique for generating a chaotic sequence uses the variable transformation and the steepest descent method as follows. By using an appropriate diffeomorphic function \( g : \mathbb{R}^n \rightarrow X \), we define \( y := g^{-1}(x) \) and \( E(y) := f(g(y)) \), and transform problem (P1) into the unconstrained optimization problem, \( \min_y E(y) \). For a rectangle constraint set \( X \), gradient model using transformation and the steepest descent method to the problem and obtain:

\[
\begin{align*}
(S1) \quad g(t + 1) &= y(t) - \alpha \nabla_y E(y(t)), \\
x(t + 1) &= g(g(t + 1)).
\end{align*}
\]

A sequence generated by (S1) converges to a local minimum which depends on its initial point, if step-size \( \alpha \) is sufficiently small. On the other hand, if \( \alpha \) is sufficiently large, it is well known that the sequence is chaotic in the sense of Marotto. Many metaheuristic methods exploit this property and search for a desirable solution. In this paper, we call this model the gradient model using transformation of variables (GT). Now, let us consider the sequence generated by GT model for the following problem:

Wood-Colville function

\[
\begin{align*}
\min \quad f(x) &= 100(2.5 - x_1^2)^2 + (1 - x_1)^2 \\
&\quad + 90(2.5 - x_2^2)^2 + (1 - x_2)^2 \\
\text{s.t.} \quad |x_1| \leq 2.5, |x_2| \leq 2.5.
\end{align*}
\]

Fig.1 shows the points generated by (S1) for this problem from some different initial points at each step-size \( \alpha \). From this, we can see that the sequences converges to two fixed points for the small \( \alpha \) and that for the sufficiently large \( \alpha \), it is chaotic, where it covers the whole feasible region.

However, even if \( \alpha \) is sufficiently large, there exist some intervals of the bifurcation parameter \( \alpha \) such that any sequence converges to several periodic points, e.g., \( \alpha \in (0.008, 0.009) \) in Fig.1, while the system is theoretically chaotic. Such intervals are called windows. What kinds of windows exist in the system mainly depends on the objective function. Thus, we should select the step-size to avoid these windows. In addition, a generated sequence often tends to overconcentrate around the boundary of the feasible region, as shown in Fig.1, which is caused by a large step-size and the diffeomorphic function for the transformation, and it does not depend on the objective function. Hence, in some cases, solutions obtained by the GT model might be undesirable.

In this paper, in order to improve the drawbacks, we propose a new chaos generator which is suitable for optimization.

4. Chaos generator with perturbation terms

Now, let us consider the following problem (P2) with sinusoidal perturbation terms instead of (P1):

\[
(P2) \quad \min f(x) + \sum_{i=1}^{n} \beta \sin(\omega x_i) \quad \text{s.t.} \quad x \in X
\]

Then, we apply a steepest descent method for problem (P2) by using an appropriate internal penalty function \( p(x) \) for the constraint set \( X \), where all local minima of the obtained objective function are interior points of \( X \). For example, for a rectangle constraint set \([0, 1]^n\), we can use the penalty function:

\[
p(x) = \sum_{i=1}^{n} c_p \tan^2(\pi(x_i - 0.5)),
\]

where \( c_p \) is an appropriate positive constant. Then, we have

\[
(S2) \quad x(t + 1) = h(x(t)) = x(t) - \alpha \left( \nabla f(x(t)) + \nabla p(x(t)) + \beta \omega \text{diag} \{ \cos(\omega x_1(t)), \ldots, \cos(\omega x_n(t)) \} \right),
\]

where \( \omega \) and \( \beta \) are a frequency and an amplitude, respectively. We call the proposed model the gradient model using original variables with the sinusoidal perturbation (GSp). In the GSp model, if \( \omega \) is set to be sufficiently large, even though \( \alpha \) and \( \beta \) is not so large, system (S2) has a larger number of fixed points than system (S1) does because of the perturbation terms, and thus most of them can be expected to be snap-back repellers as shown in Fig.2. The existence of snap-back repellers guarantees that the system generates a chaotic sequence. In addition, we derive the sufficient conditions for \( \alpha \), \( \beta \) and \( \omega \) under which system (S2) generates a chaotic sequence.

**Theorem 2** If for an interior fixed point \( x_{S2}^* \) of (S2), there exist a positive constant \( r \) satisfying \( r > 2\pi/\omega \) and a neighborhood \( N_r(x_{S2}^*) \subset X \) such that \( \nabla f(x) + \nabla p(x) = 0 \) for
all $x \in N_r(x^*_S)$. Moreover $\alpha$, $\omega$ and $\beta$ satisfy
\[ \frac{3}{2^2} \leq \alpha/\beta \omega^2, \]
then the fixed point $x^*_S$ is a snap-back repeller.

In many cases, if we select a sufficiently large $\omega$, there exist $r > 2\pi/\omega$ and a neighborhood $N_r(x^*_S) \subseteq X$ such that $\nabla f(x) + \nabla p(x) \approx 0$ for all $x \in N_r(x^*_S)$. Hence, even if $\alpha$ and $\beta$ are not so large, all assumptions of Theorem 2 can be expected to be approximately satisfied. Thus, from this theorem, we can see that the proposed model can avoid overconcentration around the boundary of the feasible region because of a small step-size. In addition, although it has more parameters than the GT model, they can be easily selected by Theorem 2.

Next, let us consider the windows appear in the GOsp model. Here, we consider the one-dimensional system, which is enough to investigate windows in system (S2). In general, about the windows, the following property is known [6]. Any window is generated by the tangent bifurcation and disappeared at the crisis. If a right-hand side of the update system (S) is approximated by a quadratic function at a stationary point $x^*$ of $f$ as follows:
\[ f(x) \sim a(\alpha)(x - x^*)^2 + b(\alpha), \]
where $\alpha$ is a bifurcation parameter. Then, the width $|\Delta \alpha|$ of a window generated by the tangent bifurcation around $x^*$ is approximately given by
\[ |\Delta \alpha| \sim \frac{2}{|a(\alpha)| b'(\alpha)}. \]

Hence, the width of the window is in inverse proportion to $a(\alpha)$. Therefore, since the curvature of right-hand side $b(x)$ of (S2) at its stationary point is large for a sufficiently large $\omega$, the width of each window can be expected to be small.

Fig. 3 shows the bifurcation diagram of (S2) for the Wood-Colville function. We can see that for a small step-size, a chaotic sequence is generated and that the width of each window is very small. Furthermore, we quantify the chaos by evaluating the Lyapunov exponent of (S2). It is well known that if the index is positive, the system is chaotic [6]. The result is shown in Fig. 4. From this figure, we can observe that under a small step-size $\alpha$, (S2) is chaotic.

5. Application to optimization

In this section, in order to compare GT with GOsp, we applied some metaheuristic methods with two generators to an optimization problem, where we used the following methods which exploit the chaotic dynamics.


This method is similar to the simulated annealing method, which uses a sufficiently large step-size $\alpha$ at early stages to search extensively for a solution in the feasible region, and then gradually decreases $\alpha$ until a sequence generated by the system converges to a local minimum.

5.2. Chaotic multi-start method: CMS [2, 3]

The CMS is a multi-start method where local searches (LS) are executed from points found by a global search (GS) making use of a chaotic sequence. In this method, the GS procedure is executed until the algorithm terminates.

Figure 2: The graph of (S2)

Figure 3: The bifurcation diagram of GOsp model

Figure 4: Lyapunov exponent of (S2)
If the GS procedure finds a promising area which can be expected to include a good local minimum, then a LS procedure is started from the point. If the solution obtained by LS is better than the tentative one, then it is updated by the obtained solution. The starting conditions of the LS procedure should be selected appropriately for each problem.

5.3. Numerical experiments

In this subsection, we apply CA or CMS with two generators GT and GOsp to an optimization problem, and verify the ability of them to find solutions. We apply them to the minimization problem for the following 10-dimensional Rastrigin function:

\[
\begin{align*}
\text{min} & \quad f_R(x) = 10n + \sum_{i=1}^{n} x_i^2 - 10 \cos(2\pi x_i) \\
\text{s.t.} & \quad |x_i| \leq 5.12, \quad i = 1, \ldots, n,
\end{align*}
\]

where the global optimum is \(x^* = (0, \ldots, 0)^T\).

First, we executed CA’s with two generators where \(\alpha\) was arithmetically decreased by \(1.2 \times 10^{-6}\) in GT and by \(5.0 \times 10^{-7}\) in GOsp, and \(\beta = 1\) and \(\omega = 100\), and the maximal number of iterations was set to be 1000 for both methods. The average objective function value obtained by each method in 50 trials is shown in Table 1. GOsp found the better solutions than GT.

| Table 1: Comparison of obtained function values for Rastrigin function |
|---------------------|---------------------|---------------------|
| Ave. func. value    | GT      | GOsp   |
| 169.9               | 19.6    |

Secondly, we compared CMS’s with two generators. Additionally, we executed the randomized algorithm (RA), in which a start point of LS was selected uniformly at random from the feasible region, instead of a chaotic trajectory. For each method, we executed system (S1) with sufficiently small \(\alpha\) as a LS procedure. Fig. 2 shows comparison of the average objective function values obtained by three methods, where the maximal number of iterations was set to be 10000 for each method. CMS with the proposed generator found the best solution among three methods. Next, we compared the number of iterations until each method find the global optimal solution. Fig. 3 shows the comparison of the average number of iterations in GS (No.of GS), starting LS’s (No.of LS), total iterations in all LS’s (No.of total ite. LS), and total iterations in the whole algorithm (No.of total ite.). GOsp requires the smallest number of iterations in terms of all indices among three methods.

From these results, we can conclude that chaotic metaheuristic methods with the proposed method can effectively find better solutions, and can find the global optimal solution within the smallest number of iterations. Those results indicate good performance of the proposed chaos generator.

| Table 3: Comparison of numbers of iterations for Rastrigin function |
|---------------------|---------------------|---------------------|
| Ave. func. value    | GT      | RA    | GOsp   |
| 2036               | 10436   | 1341 |

6. Conclusion

In this paper, we have pointed out drawbacks of the exiting chaos generator based on the steepest descent method, which has been used in many chaotic metaheuristic methods for the global optimization. In order to overcome the drawbacks, we have proposed a new chaos generator exploiting the sinusoidal perturbation terms. In addition, we have shown the sufficient conditions for parameters in the term and the step-size under which the proposed method generates a chaotic sequence. Furthermore, we have applied the chaotic annealing method or the chaotic multi-start method with the proposed generator to global optimization problems and verified the efficiency of the proposed method.

References

Metaheuristic Algorithms for Container Loading Problem
Using Grouping Objects

Yasunori Endo†, Sadaaki Miyamoto†, Koki Hanzawa‡, Yukihiro Hamasuna‡
† Department of Risk Engineering, School of Systems and Information Engineering,
University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki 305-8573, Japan
E-mail: {endo,miyamoto}@risk.tsukuba.ac.jp
‡ Master’s Program in Risk Engineering, Graduate School of Systems and Information Engineering,
University of Tsukuba, 1-1-1, Tennodai, Tsukuba, Ibaraki 305-8573, Japan

Abstract—A family of automatic container loading problems is studied and algorithms are proposed. The algorithms are constructed with metaheuristics and include flat and/or vertical loading schemes, loading efficiency, stability of loaded objects, and computational requirement. Handling groups of objects in a metaheuristic scheme is moreover considered.

1. Introduction

Container loading problems have been studied by a variety of different methods. Yet they are attracting more researchers’ attention now, mainly due to remarkable progress in distribution systems and automatization in freightage.

In this paper we consider a system for a class of automatic container loading problems. Many related studies formulate such problems as optimization of allocating boxes or objects of various shapes, e.g., [1, 2, 3]. However, knowledge on loading is considered to be more important than optimization in the sense of efficient space utilization in a container. Our approach to this problem is incorporation of various types of knowledge in a system of optimization. For this purpose we use the methodology of metaheuristics in which a simple greedy algorithm is employed to allocate objects, whereby the incorporation of different types of knowledge becomes possible.

We first describe a general framework in which an optimization system incorporating knowledge is developed. Two different optimization algorithms for allocating objects to be loaded are given.

A common sense knowledge is that, subsets of objects form natural groups which are of the same shapes or their destinations are the same. The former is the way of grouping based on physical attributions of each object, e.g., weight, capacity, and the latter is based on non-physical attributions of each object.

Hence a family of search algorithms taking the grouping into account is studied, in which neighborhoods of groups are employed. Moreover mixed strategies of an ordinary and group-based searches are proposed.

2. Problem Description

As shown above, we formulate the loading problem as that of metaheuristic optimization [5, 6] in which knowledge incorporation is done through several methods.

First, formulation knowledge is incorporated already in the metaheuristics and the choice of greedy algorithms, since a greedy algorithm is a simple method that are flexible to knowledge incorporation. Second, knowledge base should be prepared and converted into algorithms of the system. As shown above, they should be represented as a part of criteria in a multiobjective optimization, constraints, and parameter values.

These observations lead us to the following basic framework of general loading problems that consists of a triplet: \( K, M, T \) where

1. \( K \) is a knowledge base in which different types of knowledge is included;
2. \( M \) is a metaheuristic optimization system of which the substructure is defined below;
3. \( T \) is a transformation mechanism whereby elements in \( K \) is transformed into working knowledge in \( M \).

Among the three components, the metaheuristic optimization system should be especially remarked, since the metaheuristic structure of the system controls the other two.

The system of metaheuristic optimization has eight components:

\[ M = (O, SO, N, U, C, L, HA, lc) \]

The first component is the set of objects to be loaded: \( O = \{o_1, \ldots, o_N\} \). Moreover \( SO \) is all sequences of the objects

\[ SO = \{x = (o_{j_1}, \ldots, o_{j_N})\}, \]

where \((j_1, \ldots, j_N)\) runs all permutations of \((1, \ldots, N)\). The components \( U \) and \( L \) imply the upper and lower stages. In the lower stage a certain heuristic algorithm \( HA \) is chosen to allocate each object \( o_i \) in a given order \( x \) in \( SO \).

In the upper stage \( U \) the order of the objects is altered using a metaheuristic method such as the local search, the simulated annealing [4], the tabu search, or the genetic algorithm.

A solution is represented by the sequence \( x \) of objects. Together with an allocation subalgorithm \( HA \), the locations of the objects are determined. By using a metaheuristic method, the solutions are iteratively optimized. In a metaheuristic method, the concept of neighborhood of a solution is used. We assume a
neighborhood of $x$ is denoted by $N(x)$. A typical example of $N(x)$ is the exchange: an element $y \in N(x)$ is obtained by exchanging two objects in the sequence $x$.

Generally an allocation subalgorithm uses a heuristic local criterion $l(c(\cdot))$ for finding a location of an object $o_i$, which is dependent on particular implementation of $HA$. On the other hand, the optimization in the metaheuristic level $U$ uses another global criterion $C$. These two criteria depend strongly on individual application and the algorithm, as we will see later.

There are many choices for the algorithm $HA$ in $L$ and many different metaheuristic schemes in $U$. Moreover the choices are independent between $U$ and $L$.

As noted before, knowledge is incorporated in both $U$ and $L$. The global criterion $C$ includes a weighted sum of soft constraints while the local criterion also can. The greedy algorithm includes constraints when allocating objects.

3. Allocation Algorithms

In this section we propose algorithms for allocating objects to be loaded and optimization schemes. We assume that all objects and the container are rectangular boxes. The purpose of the algorithms is to maximize the volume ratio of the container as follows:

$$ R = \frac{\sum_j v_j}{C} \rightarrow \max $$

Here, $C$ and $v_j$ are respectively the volume of the container and that of $o_j$ which are allocated in the container.

3.1. Loading strategies

In this paper, we consider two strategies of loading. One is vertical loading and the other is flat loading.

3.2. Constraints

3.2.1. Constraint of a container

All objects should be allocated inside a container so that no object is stuck out of the container.

3.2.2. Rotation of objects

Rotation of an object is sometimes prohibited and in other cases free. We consider two ways to rotate each object. In one way any side of an object can be the bottom, in the other way any objects can not be turned upside down.

3.2.3. Stability of objects

The measure of stability $S$ of an object which is allocated onto other objects is defined as follows:

$$ S = \frac{\text{contact space between upper object and lower objects}}{\text{underside space of upper object}} $$

We called $S$ the ratio of contact between the upper object and lower objects. The range of $S$ is $0.5 < S \leq 1$. $S \neq 1$ means that a part of the underside of the upper object is not completely contacted onto the lower objects.

3.3. Evaluation of allocation position of objects

We evaluate the following criteria in view of stability to finalize allocation positions of objects.

1. Protruding ratio of objects
2. Degree of the ‘brick stacking’ pattern
3. The average number of the brick stacking patterns

3.3.1. Protruding ratio

The allocation with $S \neq 1$ is not desirable in general. Hence, we define the ratio of protruding $O_p$ as follows:

$$ O_p = 100 \times \frac{\text{total sum of protruding underside space of objects}}{\text{total sum of underside space of the allocated objects} \times (1 - S)} $$

3.3.2. Degree of brick stacking pattern

The ‘brick stacking’ pattern means that an upper object should be allocated regularly on the boundary of lower two objects. Such allocation is obviously robust to vibration in transportation.

We define the degree of brick stacking pattern $B_p$ by the following, where ‘n.’ implies ‘numbers’ and the lowest objects are on the floor of the container.

$$ B_p = 100 \times \frac{n. \text{ of objects allocated in brick stacking pattern}}{n. \text{ of all allocated objects} - n. \text{ of the lowest allocated objects}} $$

3.3.3. Average number of objects in brick stacking pattern

We introduce the average number of objects in the brick stacking pattern $L_{ave}$ as follows:

$$ L_{ave} = \frac{n. \text{ of objects right under each object}}{n. \text{ of allocated objects}} $$

Here we assume that the number of objects which are right under the lowest object is 1.

3.4. Allocation Subalgorithm $HA$

In this section, we describe two allocation subalgorithms in $L$ to find the location of an object $o_i$. These subalgorithms are both based on a simple greedy algorithm, but management of data structure is different. In the following subalgorithms stability criteria in the last section are compared against predetermined thresholds.
3.4.1. Cell division management

A method to manage objects in the container by cell division (cf. [7]), which means that a container is divided into small cells and when an object is allocated, the space occupied by the object is marked ‘1’ and those not occupied is marked ‘0’, whereby management of free space is simple and flexible. The method is called here ‘cell division’ management.

3.4.2. Vertex management

Second algorithm uses vertices of objects to manage the space [8]. The method is called ‘vertex management’.

3.5. Metaheuristic Schemes

We use a local search and the simulated annealing as metaheuristic schemes in $U$. A local search and the simulated annealing employ search in a neighborhood and an appropriate definition of a neighborhood $N(x)$ of an object $x$. Notice that a solution $x$ is an order of all objects. In such a case a natural neighborhood is obtained by exchanging two objects. That is, for

$$x = (o_1, \ldots, o_{i-1}, o_i, o_{i+1}, \ldots, o_{j-1}, o_j, o_{j+1}, \ldots, x_N)$$

We define

$$N(x) = \{x \mid x = (o_1, \ldots, o_{i-1}, o_i, o_{i+1}, \ldots, o_{j-1}, o_j, o_{j+1}, \ldots, x_N), \forall i, j\} \quad (1)$$

4. Object Groups

In this section, we consider the way to have grouping of the objects. As noted earlier, we have natural groupings of the objects:

- The groups of objects are determined based on physical attributions of each object, e.g., weight, capacity.
- The groups of objects are determined based on non-physical attributions of each object, e.g., destinations.

In both cases, a group of objects should be placed closely, which is common sense knowledge in most loading problems. Now, we discuss the former grouping.

4.1. Grouping using Clustering

Let $a_{ik}$ be $k$-th attribution of an object $o_i$ and $a_{i} = (a_{i1}, \ldots, a_{iP})$ be an attribution vector of $o_i$. We can show the following examples of the attributions:

- weight
- capacity
- sum of each edge

The flow of having groups is that, first give the number of groups and choose a clustering method, next group the objects using the chosen method. One of the simplest methods is as follows:

1. Give the number of groups. Let the number be $M$.
2. Calculate the attribution of each object. For example, the weight or capacity can be regarded as the attribution. In convenience, the attribution is not a vector but a scalar.
3. Sort the objects in descending order of the value of attribution. Let the sequence of objects be $(o_1, \ldots, o_N)$ and $a_i$ be the attribution of $o_i$.
4. Calculate the value $d_{i,i+1} = a_i - a_{i+1}$.
5. Find the $M - 1$ largest values of $d_{i,i+1}$. Let the $i$ be $i_k$. Note that number of $i_k$ is $M - 1$.
6. Group the object as follows:

$$\{[a_1, \ldots, a_{i_k}, a_{i_k+1}, \ldots, a_{i_l}, a_{i_l+1}, \ldots, a_N]\}$$

Of course, we can also use other clustering algorithms, e.g., fuzzy c-means, hierarchical clustering.

4.2. Reflection of Grouping into Metaheuristic Optimization

To reflect such grouping into the metaheuristic optimization, a natural strategy is to alter a neighborhood from (1) into another type of neighborhoods based on the groups.

Let a group $G_k$ is a sequence of objects:

$$G_k = (o_{i+1}, \ldots, o_j).$$

For later use, we denote the set of objects in $G_k$ by

$$[G_k] = \{o_{i+1}, \ldots, o_j\}$$

and $|G_k|$ is the number of objects in $G_k$.

It should be noticed that frequently the sequence itself is not essential but the set $[G_k]$ is generally important. For two sequence $G$ and $G'$, if $G \neq G'$ and $|G| = |G'|$, then $G$ and $G'$ have actually the same object set.

Assume the solution $x$ is separated into sub-sequences:

$$x = (o_1, \ldots, o_{i_1}, o_{i_1+1}, \ldots, o_{i_2}, o_{i_2+1}, \ldots, o_N)$$

and each subsequence forms a group:

$$x = (o_1, \ldots, o_{i_1}, o_{i_1+1}, \ldots, o_{i_2}, o_{i_2+1}, \ldots, o_N)$$

$$\Downarrow$$

$$x = (G_1|G_2|\ldots|G_k|G_{k+1}|\ldots|G_l|G_{l+1}|\ldots|G_M)$$

Thus we have a sequence of groups instead of the original sequence of individual objects.
For \( x = (G_1|G_2|\ldots|G_k|G_{k+1}|\ldots|G_l|G_{l+1}|\ldots|G_M) \), we can define a new neighborhood \( N_G(x) \) by groupwise exchange:

\[
N_G(x) = \{ x | x = (G_1|G_2|\ldots|G_k|G_{k+1}|
\]
\[
\ldots|G_l|G_{l+1}|\ldots|G_M), \forall k, l \}
\]

Handling an object sequence by groups has two effects:

1. Objects in a group become a subsequence and hence placed closely in a container. This means that, e.g., objects in a group of the same destination are placed as a cluster and handling at a freight station becomes easier.

2. Search space becomes narrower by using the groups and hence the greedy subalgorithm is more efficient.

### 4.3. Mixed search

Mixing an ordinary search by individual objects using \( N(x) \), which is also called a standard search (abbreviated as \( S \)), and the groupwise search using \( N_G(x) \) (abbreviated as \( G \)) is studied. Notice that using a standard search \( x \) does not have a group as a subsequence in general, while a groupwise search requires those subsequences. Let us assume that \( x = (G_1|G_2|\ldots|G_k|G_{k+1}|\ldots|G_l|G_{l+1}|\ldots|G_M) \). By using a standard search, \( x \) is randomized into \( x' \) in which \( G_i \) is scattered. For example \( G_1 = (o_1, o_2, \ldots, o_l) \) and we may have \( x' = (o_1, \ldots, o_2) \). Suppose we wish to perform a groupwise search for \( x' \).

#### Recovery of groups

To recover the groups from a randomized \( x' \) is necessary. A simple method is to group the objects into original groups in which the order of the objects keeps the order of \( x' \). We show a simple example as follows:

1. Let an initial set be

\[
x = (o_1, o_2, o_3, o_4, o_5, o_6, o_7, o_8, o_9)
\]

which has already been sorted in descending order of the value of attribution of each object.

2. As a result of clustering, \( x \) is divided in three groups as follows:

\[
x = (o_1, o_2, o_3|o_4, o_5, o_6|o_7, o_8, o_9)
\]

\[
\Downarrow
\]

\[
x = (G_1|G_2|G_3)
\]

3. As a result of standard search after groupwise search, we get the following sequence:

\[
x = (o_4, o_8, o_3, o_7, o_9, o_6, o_1, o_2, o_5)
\]

4. As a result of recovery of groups, we can get the sequence:

\[
x = (o_4, o_6, o_5, o_8, o_7, o_9|o_3, o_1, o_2)
\]

### 5. Conclusion

In this paper, two algorithms of cell division and vertex management have been proposed. We moreover have proposed searches based on object groups. The way to group the object using clustering have been proposed and mixed algorithms \( S-G-S \) and \( G-S-G \) have been discussed. Generally such group handling is considered to be important as a common sense type of knowledge. Although we have tested a number of examples, we cannot show a definite advantage of the group handling. However, we have shown such group handling algorithms are at least as good as an ordinary metaheuristic algorithm.

Future study include test on actual sets of objects at freight stations and evaluation of the proposed algorithms using more extensive criteria. Expert evaluation by experts in actual freight stations is also necessary.

### Acknowledgments

We thank Dr. Shigeyuki Takahara of Kagawa Prefectural Industrial Technology Center, Mr. Kunimitsu Urayama of the Sanukite Co. Ltd., and Mr. Daisuke Kobayashi of the MTI Co. Ltd. for their helpful suggestions. This study has been supported by the Program for Promoting Fundamental Transport Technology Research from the Japan Railway Construction, Transport and Technology Agency (JRTT), Project Number 2004-01.

### References


On-line Reference Inputs Management for Constrained Servo Systems

Y. Ohta†, T. Taguchi† and T. Yamaguchi†
†School of Engineering, Kobe University
1-1 Rokkodai, Nada, Kobe 657-8501, Japan
Email: ohta@cs.kobe-u.ac.jp

Abstract—In this paper, we investigate a method to control servo systems using reference governor and the BB method. To switch reference signals by the reference governor, a maximal admissible sets (MAS) corresponding to a reference signal, which includes the current state, must be determined on line. In this paper, we propose an efficient method to do this. This method uses the BB method and a balanced binary tree concept. The BB method is also used to compute MASs.

1. Introduction

For almost all practical control systems, we need to take into account the constraints on state and/or control input caused by amplitude limitation of state variables, saturation property of actuators and so on. If we ignore these constraints, then the real performance of the system degrade because of the wind-up phenomena, or in worst cases the control system becomes unstable. In these respect, extensive researches have been done to cope with such constraints, and predictive reference management methods [1], [2], a reference signal shaping method [3], and methods using reference governor (RG) [4]–[8] are proposed. It was said that the method using RG could not improve the original output response. However, in [9], it has been shown that the original output response could be improved by adopting both RG and an outer feedback loop which is designed by using LMI.

In this paper, we study a method adopting RG and an outer feedback loop. The first issue to utilize the advantage of the method is the computation of the maximal admissible sets (MASs) for constants reference signals. The second issue is the fast determination that MASs which include the current state of the servo system. The first issue is rather easy since computing MASs is off-line task. On the other hand, the second issue is very crucial to use RG to manage reference input, because it is on-line task and should be done in a given sampling period. We will propose a method using Beneath-Beyond (BB) method [10], [11], which is a method to solve a dynamic convex hull problem.

Notation. For a vector $\xi$, $[\xi]_p$ denotes the $p$-th element of $\xi$. For integers $p$ and $q$ such that $p \leq q$, $[p \cdots q]$ denote the set $\{p, p+1, \ldots, q\}$. For a set $\Omega$, $\text{int} \ \Omega$ is the set of interior points of $\Omega$. For a polytope $\mathcal{Z}$, facet $\mathcal{Z}$, node $\mathcal{Z}$, and $\text{node} \ \mathcal{Z}$ are a set of all facets of $\mathcal{Z}$, a set of all nodes of $\mathcal{Z}$, the cardinality of facet $\mathcal{Z}$, and the cardinality of node $\mathcal{Z}$, respectively.

2. Problem Statement

Let us consider a discrete-time servo system given by

$$\begin{align*}
x[k+1] &= Ax[k] + Br[k], \quad x[0] = x_0 \\
y[k] &= Cx[k],
\end{align*}$$

where $k \in \mathbb{Z}_+ = \{0, 1, 2, \ldots\}$, $x \in \mathbb{R}^n$ is the closed loop state, $y \in \mathbb{R}^m$ is the output to be controlled, and $r \in \mathcal{R}\subseteq \mathbb{R}$, where $\mathcal{R}$ is the given connected set to which constant reference signals belong. We assume (1) is a stable servo system, and, hence, the pair $(A, B)$ is controllable; $A$ is a stable matrix; and for any $r \in \mathbb{R}^m$ there exists a unique $\hat{x}(\hat{r})$ such that

$$\hat{x}(\hat{r}) = A\hat{x}(\hat{r}) + B\hat{r}, \quad C\hat{x}(\hat{r}) = \hat{r}. \quad (2)$$

We require constraints on state and control are satisfied. The constraints are given by

$$z[k] = Lx[k] + Dr[k] \in \mathcal{Z} \subseteq \mathbb{R}^{N_c}, \quad \forall k \in \mathcal{N}, \quad (3)$$

where $\mathcal{N} = [0 \cdots N] \subseteq \mathbb{Z}_+$ and $\mathcal{Z} \subseteq \mathbb{R}^{N_c}$ is a polyhedral set given by

$$\mathcal{Z} = \{ z = [z_1 \ z_2 \ \cdots \ z_{N_c}]^T \mid |z_i| \leq 1, \ i \in \mathcal{N}_c \}, \quad (4)$$

where $\mathcal{N}_c = \{1, 2, \ldots, N_c\}$.

For a given $r_j \in \mathcal{R}$, let us define

$$\Omega_k(r_j) = \{ x_0 : (Lx[k;0,x_0,r_j] + Dr_j) \in \mathcal{Z}, \ k \in [0 \cdots \hat{k}] \}, \quad (5)$$

where $x[k;0,x_0,r_j]$ is the solution of (1) satisfying $x[0] = x_0$ and the input $r[k] \equiv r_j$. The MAS corresponding to a constant reference input $r_j$ is denoted by $\Omega_\infty(r_j)$ and defined by

$$\Omega_\infty(r_j) = \lim_{k \to \infty} \Omega_k(r_j). \quad (6)$$

Throughout the paper, we assume that

$$\forall \hat{r} \in \mathcal{R}, \ (L\hat{x}(\hat{r}) + D\hat{r}) \in \text{int} \ \mathcal{Z}, \quad (7)$$

and

$$\forall r \in \mathcal{R}, \ \Omega_\infty(r) \neq \emptyset, \ \hat{x}(r) \in \text{int} \ \Omega_\infty(r). \quad (8)$$
We divide $\mathbf{R}$ into smaller sets $\{\mathcal{R}_j\}$ and choose a point $r_j \in \mathcal{R}_j$ for each $j$. Suppose that we computed $\{\Omega_\infty(r_j)\}$ and that 

$$
\hat{x}(r_j) \in \text{int } \Omega_\infty(r_{j'}) \quad \forall j' \in J(j),
$$

(9)

where $J(j)$ is the set of indexes of $\mathcal{R}_j$’s which are adjacent to $\mathcal{R}_j$. Now suppose that an initial state $x_0$ and a constant reference command $\ell^*$ is given and let $x_0 \in \Omega_\infty(r_j)$ and $r^* \in \mathcal{R}_j$. Consider a path connecting $r_h$ and $r_{j'}$, and suppose that it passes $\mathcal{R}_h, \mathcal{R}_{j'}, \ldots, \mathcal{R}_{j'}$. Then, RG manages $[r[k]]$ in the following way: For each $k$, determine the largest $\ell(k) \in [0 \cdots N]$ such that $x[k] \in \Omega_\infty(r_{j_0})$, then set $r[k] = r_{j_0}$ so that the following relations hold.

$$r[k] = \begin{cases} 
  r_{j_0}, & x[k] \in \Omega_\infty(r_{j_0}), \\
  r', & x[k] \in \Omega_\infty(r'), \quad k \geq k_N.
\end{cases}
$$

(10)

(11)

When a polytope $\mathcal{Z}$ is given by

$$\mathcal{Z} = [h_i^T z \leq 1, \ i \in [1 \cdots m_Z]],
$$

(12)

we have

$$
\Omega_k(r_j) = \begin{cases} 
  \Omega_k'(r_j) + \hat{x}(r_j), & e_0 \in \Omega_k'(r_j), \\
  \Omega_k'(r_j), & e_0 \notin \Omega_k'(r_j)
\end{cases}
$$

(13)

(14)

(15)

where $\hat{z}(r_j) = L\hat{x}(r_j) + Dr_j$, and

$$
\hat{h}_i^T = \frac{1}{1 - h_i^T \hat{z}(r_j)} \hat{h}_i^T LA^k.
$$

(16)

Note that $0 \in \text{int } \Omega_k'(r_j)$ since $\hat{x}(r_j) \in \text{int } \Omega_k'(r_j)$ by (8).

3. Computation of MAS and the BB Method

In the following, we drop $(r_j)$ when it is obvious from the context.

3.1. Algorithm for computation of MAS

The following result can be obtained easily.

**Lemma 1** Let

$$
\Omega_k^D = \text{co } \{\hat{h}_i, \ i \in [1 \cdots m_Z], \ k \in [0 \cdots \hat{k}]\},
$$

(17)

node $\Omega_k^D = \{\hat{h}_{\ell(k),k}; \ell \in [1 \cdots m_k]\}.
$$

(18)

Then, we have

$$\Omega_k^D = \{x : \hat{h}_{\ell(k),k}^T x \leq 1, \quad \forall \ell \in [1 \cdots m_k]\}.
$$

(19)

**Lemma 2** Let $\tilde{\Omega}_k^{D+1}$ be $\text{co } \{\text{node } \Omega_k^D; \quad \hat{h}_{i,k+1}, \ i \in [1 \cdots \hat{k}]\}$. If $\hat{h}_{i,1,k+1} \in \Omega_k^D$ then we have

$$
\hat{h}_{i+1,k+j} \in \tilde{\Omega}_k^{D+1} \quad \forall j \geq 1.
$$

(20)

Based on Lemma 2, we propose an algorithm to compute $\Omega_\infty^D$.

**Algorithm**

1. Let $Q$ be an empty queue, and let $\Omega_k^D$ be a sufficiently small polytope such that $0 \in \text{int } \Omega_k^D$.
2. $\text{co } (\Omega_k^D \cup \hat{h}_0); \quad \text{push}(Q, \hat{h}_0); \quad \text{co } (\Omega_k^D \cup \hat{h}_0); \quad \text{push}(Q, \hat{h}_0); \ldots; \quad \text{co } (\Omega_k^D \cup \hat{h}_0); \quad \text{push}(Q, \hat{h}_k^0);
3. while $(Q \neq \emptyset)$
4. $h := \text{pop}(Q); \quad \hat{h} := A^T h$;
5. if $(\hat{h} \notin \Omega_k^D)$ then begin
6. $\Omega_k^D := \text{co } (\Omega_k^D \cup \hat{h}); \quad \text{push}(Q, \hat{h});$
end;

end.

In the above, $\text{push}(Q, h)$ means that add $h$ at the end of $Q$, and $\text{pop}(Q)$ returns the first element of $Q$ and the element is removed from $Q$.

**Remark 1** By the assumption, $A^k \to 0$, $k \to \infty$, and, hence, $\hat{h}_k \to 0$, $k \to \infty$. Therefore, it is obvious that the above algorithm will stop anyway.

**Remark 2** The algorithm propose here coincides with that proposed in [12]. However, the implementation of the algorithm is different as we show in below.

### 3.2. The BB Method

Suppose that $0 \in \text{int } \Omega_k^D$. Equation (19) means that $m_k = |\text{facet } \Omega_k^D|$, and that $\hat{h}_{\ell(k),k}$ is a normal vector of $\mathcal{F}_k \in \text{facet } \Omega_k^D$. Note that $\hat{h}_{\ell(k),k}$ is normalized in the sense that it satisfies

$$
\hat{h}_{\ell(k),k}^T x = 1 \quad \forall x \in \mathcal{F}_k.
$$

(21)

Then, polytopes $\Omega_k^D$ and $\Omega_k^D$ are dual, and there is 1-1 correspondence between nodes’ (normalized normal vectors of facets, respectively) of $\Omega_k^D$ and normalized normal vectors of facets (nodes, respectively) of $\Omega_k^D$. For example, let us consider a cube $\Omega_1^D$ shown in Fig. 1, where node $\Omega_1^D = \{x_1, x_2 \cdots, x_8\}$. By 1234, let us denote a facet whose nodes are $x_1, x_2, x_3, x_4$. Then, facet $\Omega_1^D = \{F_1: 1234 \text{ (the facet } F_2: 1265, F_5: 3478, F_4: 5678, F_3: 1485, F_6: 2376\}$. The node $h_i$ of the octahedron $\Omega_k^D$ is the normalized normal vector of $\mathcal{F}_i$, where

$$
\begin{bmatrix}
  x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\
-1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\
-1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\
-1 & -1 & 1 & -1 & -1 & 1 & 1 & -1 \\
0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
-1 & 0 & 0 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}.
$$
In the BB method proposed in [11], a polytope keeps the data of nodes, facets, sub-facets and the inclusion relations of them (for example, see Fig. 2).

![Fig. 1 A cube $\Omega'_k$ and its dual polytope $\Omega^D_k$.](image)

**Fig. 1** A cube $\Omega'_k$ and its dual polytope $\Omega^D_k$.

Let node $\Omega'_k = \{x_j, \ j \in [1 \cdots n]\}$. Then,

$$\hat{h} \notin \Omega^D \iff \exists j \in [1 \cdots n]: \ x_j^T \hat{h} > 1,$$

and, hence, $\hat{h} \notin \Omega^D$ holds if and only if (22) holds. This can be done very easily and very much faster than the method solving Linear Programming Problem (LP)[12]. Note that in the algorithm, we need to compute $\cos (\Omega^D \cup \hat{A}^T \hat{h})$ at Step 6 when $\hat{h} \notin \Omega^D$. This is additional task as long as we confine ourselves to determine $\hat{h} \notin \Omega^D$ or not. However, this removes unnecessary constrains, and it is useful to reduce the total computing cost.

Computation $\cos (\Omega^D \cup \hat{A}^T \hat{h})$ consists of 2 steps: The first step is to find a facet such that $x_i^T \hat{h} > 1$, where $x_i$ is the normalized normal vector of the facet. The second step is renewing the data of $\Omega^D$. The first step is done by $O(n)$ computing cost, where $n = |\text{facet } \Omega^D|$. On the other hand, the second step is usually done by $O(m^2)$ $\sim O(m^3)$ computing cost[11], where $m = |\text{node } \Omega^D|$.

### 3.3. Efficient method to examine $\hat{e} \in \Omega'_\infty$ or not

Suppose that we have $\Omega^D_\infty$. Let node $\Omega^D_\infty = (h_1)^m \cup$ and let $\Omega'_\infty$ be the dual polytope of $\Omega^D_\infty$. To apply the method based on RG, we need to determine whether $\hat{e} = \hat{x} \in X'_\infty$ or not on line. In this section, we propose a method to reduce the computing time to do this. By a naive method, it need $m$ times computation of the inner product since

$$\hat{e} \in \Omega'_\infty \iff \forall \ell \in [1 \cdots m]: \ \hat{h}_\ell^T \hat{e} \leq 1.$$

Let node $\Omega'_\infty = \{x_i\}^m_i$, where $x_i$ is the normalized normal vector of a facet of $\Omega^D_\infty$.

Let $\mathcal{X} = \mathcal{X} - \hat{x}$, and we divide $\mathcal{X}$ into $(\mathcal{X}_q, q \in [1 \cdots N_q])$ by hyperplanes $\{H_p, p \in [1 \cdots N_h]\}$. Suppose that $\mathcal{H}_p = \{x: h_p^T x = 1\}$, and, for each $\mathcal{X}_q$, define $\xi_q \in [0, 1]^m$ by

$$[\xi_q]_p = 1 \iff h_p^T x' \geq 0 \wedge x' \in \mathcal{X}_q.$$

![Fig. 2 The data structure of the BB method[11]: Nodes, facets, sub-facets and the inclusion relation of them.](image)

**Fig. 2** The data structure of the BB method[11]: Nodes, facets, sub-facets and the inclusion relation of them.

For example, in Fig. 3, $N_h = 3$, $N_X = 6$, $\xi_1 = [0, 0, 1]^T$, $\xi_2 = [0, 0, 0]^T$, $\xi_3 = [0, 1, 0]^T$, $\xi_4 = [1, 1, 0]^T$, $\xi_5 = [1, 1, 1]^T$, and $\xi_6 = [0, 1, 1]^T$. We introduce lexicographic order and construct a balanced binary tree $T$. Any node of $T$ corresponds to a region $\Omega'_\infty$ and has the data of normalized normal vectors of facets which includes nodes in $\Omega'_\infty$.

For example, the node corresponding to $\mathcal{X}_h$ has data of $h_3$ and $h_11$. When $\mathcal{X}_h$ includes no node of $\Omega'_\infty$, the corresponding node of $T$ has the data of $h_i$ of $\mathcal{F}_i$, where $\mathcal{F}_i$ is the facet such that

$$\alpha h_i^T x_q \geq \alpha h_i^T x_q, \ \forall \ell, \ \alpha h_i^T x_q > 1,$$

where $\alpha$ is a sufficiently large number and $x_q$ is an arbitrary given point in int $\mathcal{X}_h$. For example, the node corresponding to $\mathcal{X}_h$ has the data of $h_3$.

When $\hat{e}$ is given, we compute

$$[\xi(\hat{e})]_p = \begin{cases} 1, & n_{\hat{e}}^T \hat{e} \geq 0 \\ 0, & n_{\hat{e}}^T \hat{e} < 0 \end{cases}$$

and search $\hat{e}$ such that $\xi(\hat{e}) = \xi(\hat{e})$. For any $h_i$ stored in the node corresponding to $\mathcal{X}_h$, we examine $h_i^T \hat{e} > 1$ or not. If $h_i^T \hat{e} \leq 1$ for all $\ell$, then $\hat{e} \in \Omega'_\infty$. For example, when $\hat{e} \in \mathcal{X}_h$, $\xi(\hat{e}) = [0, 0, 0]^T$ and $\hat{e} \in \Omega'_\infty$ if $h_i^T \hat{e} \leq 1$.

### 4. An Example

Consider a position servo system shown in [8]. The closed system is transformed into discrete-time by zero-order hold assuming a sampling period of $T_s = 10$ [ms]. We use the "outer" feedback/feedforward system proposed in [9]. In Figures 4 and 5, simulation results are shown. In which, $\tilde{r}(t) = \hat{r}[k], \ r(t) = r[k], \ t \in [kT_s, kT_s + T_s)$, and
magenta lines are responses when we use MAS, while blue lines shows responses when we apply the reference input shaping method [3].

When we use MAS, computing time of a naive method when ̃ is located inside of Ω (r1), the proposed method and a naive method need about 2.66[μS] and 4.50[μS], respectively, to determine ̃ is inside of Ω (r1). When ̃ is located at the outside of Ω (r1), the proposed method and a naive method need 0.83 ~ 1.0[μS] and 0.16 ~ 1.3[μS], respectively, to determine ̃ is inside of Ω (r1). Therefore, we can see that our method is more effective.

![Graph](image)

Fig. 4 Response when x0 = 0.

![Graph](image)

Fig. 5 Response when x0 = [1.1496 0 0]T.

5. Conclusion

In this paper, we study the reference input management using RG. We proposed to use the BB method for the construction of MAS. This method can reduce the total computing cost. Moreover, we propose to construct balanced binary trees to reduce the computing time to examine ̃ ∈ Ω or not. This method reduce the computing time about 1/2 ~ 1 than a naive method. The proposed method is much more effective than the naive method when MASs are given huge number of constraints.

References


A New Absolute Stability Criterion for 2-D Single Variable Lur’e systems

Roisin Duignan† and Paul F. Curran†
†School of Electrical, Electronic and Mechanical Engineering, University College Dublin, Belfield, Dublin 4, Ireland
Email: roisin.duignan@ee.ucd.ie, paul.curran@ucd.ie

Abstract—A new absolute stability criterion for 2-D single variable Lur’e systems is presented. This criterion is better than the well-known circle criterion and the 45° criterion. Sufficient Conditions for the existence of a piecewise quadratic common Liapunov function for 2-D single variable Lur’e systems are stated in terms of Linear Matrix Inequalities (LMI’s).

1. Introduction

One of the main factors motivating the study of switched systems is the rapidly developing area of switching control. A switched control system is a type of hybrid dynamical system consisting of continuous and/or discrete time processes interfaced with some decision making process, which estimates the current active environment and selects the appropriate controller. Switched systems have many practical applications in both electrical and mechanical control systems. Applications include power systems, power electronics, air traffic control, aircrafts and satellites to name but a few. The technique of switching among different controllers or switching into different states can provide a significant improvement in performance compared to that of a fixed controller.

The type of dynamical systems that are of interest here are 2-D single variable Lur’e systems of the form

\[
\dot{x} = Ax + f(x,t)bc^T x
\]

where, \(A \in \mathbb{R}^{2\times 2}\), \(f(x,t)\) is a non negative scalar nonlinearity, \(b, c \in \mathbb{R}^2\) are linearly independent constant vectors of the form \(c^T = \begin{bmatrix} z & 1 \end{bmatrix}, z \in \mathbb{R}^{2-1}\) and \(b^T = \begin{bmatrix} \tilde{o} & 1 \end{bmatrix}\), \(\tilde{o} \in \mathbb{R}^{2-1}\), \((A,b)\) is completely controllable.

Many frequency domain stability criteria have been derived for Lur’e systems. Of particular interest are those developed using a Liapunov [4] approach, for example the Popov Criterion [5] and the Kalman-Yakubovich-Meyer Lemma, [6],[7],[8]. However these criteria only apply to time-invariant systems. One criterion of particular importance is the Circle Criterion [9] since it is an explicit time-varying frequency domain condition. If the circle criterion can establish stability, it can be shown that this is equivalent to the existence of a common quadratic Liapunov function for the system, [10].

Given this, in recent years much research on the stability of Lur’e systems has focused on the approach of constructing a common Liapunov function. Molchanov and Pyatnitski [11] established that for absolute stability of time-varying Lur’e systems it is necessary and sufficient that there exist a common Liapunov function. They also ascertained that this common Liapunov function can belong to one of four classes - common convex Liapunov functions, common homogeneous Liapunov functions, common piecewise quadratic Liapunov functions and common unic Liapunov functions. Johansson and Rantzer [12] have developed a computational approach to establishing the existence of one of these classes, common piecewise quadratic Liapunov functions, by formulating the search as a convex optimisation problem in terms of linear matrix inequalities (LMI’s). Polanski [13] has investigated ways of explicitly establishing another class, unic Liapunov Functions.

Our investigations also focus on finding common piecewise quadratic Liapunov functions, the motivation being to establish simple analytically tractable absolute stability criteria for Lur’e Systems, that improve upon and generalise the circle criterion. Since the circle criterion is equivalent to the existence of a one piece common piecewise quadratic Liapunov function, a logical step in extending the circle criterion may be to establish necessary and sufficient conditions for the existence of a two piece common piecewise quadratic Liapunov function. For class (1), there is evidence to support this approach. Previous work has shown that for a common piecewise quadratic Liapunov function to exist one only has to determine that the root locus of \(G(s)\) lie inside a specific sector of the complex plane known as the 45° region [14]. This stability criterion is known as the 45° criterion. In this paper a new absolute stability criterion for class (1) is presented. It is shown that this new criterion improves on the circle criterion and the 45° criterion for this class. Sufficient conditions for the existence of a piecewise quadratic common Liapunov function for 2-D Lur’e systems are stated in terms of LMI’s.

2. Stability Conditions for 2-D Lur’e Systems

In the following section stability results concerning nonlinear time varying single variable Lur’e systems of the
form (1) are presented.
Consider class (1)
\[
\dot{x} = \begin{bmatrix} 0 & 1 \\ -\beta & -\alpha \end{bmatrix} x - f(x,t) \begin{bmatrix} 0 \\ 1 \end{bmatrix} z
\]
where Aizerman’s [15] well-known necessary condition for absolute stability holds if \(\alpha > 0, \beta > 0, z \geq 0\). A sufficient condition for the absolute stability of class (1), is the existence of a common Liapunov function \(V(x)\) of the form
\[
V(x) = \begin{cases} x^T P_1 x & \text{for } (b^T x)(c^T x) > 0 \\ x^T P_2 x & \text{for } (b^T x)(c^T x) < 0 \end{cases}
\]
where \(x^T P_1 x = x^T P_2 x\) for \((b^T x)(c^T x) = 0\), \(P_1 = P_1^T \in \mathbb{R}^{2\times 2}\) and \(P_2 = P_2^T \in \mathbb{R}^{2\times 2}\).

### 2.1. Lemma 1

There exists a common Liapunov function \(V(x)\) of the form (2) for class (1) if there exist scalars
\[
-1 \leq \tau \leq 1, \quad -1 \leq \eta \leq 1 \quad q, \gamma \geq 0
\]
and a real symmetric matrix \(P_0 \in \mathbb{R}^{2\times 2} > 0\) such that,
\[
A^T P_0 + P_0 A + N \pm M_q \leq 0
\]
where
\[
P_0 b = \tau b + (\gamma + 1) c
\]
and \(A\) is stable.

The proof of Lemma 1 is given in the appendix.

### 2.2. Theorem 1

There exists a common Liapunov function \(V(x)\) of the form (2) for class (1) if
(i) \(\alpha > 0, \beta > 0, z \geq 0\) and \(0 < z \leq 2\alpha\).

The Circle Criterion states that class (1) is absolutely stable if \(\text{Re}[G(j\omega)] \geq 0\) for all \(\omega \geq 0\) where,
\[
G(j\omega) = \frac{j\omega + z}{-\omega^2 + j\omega\alpha + \beta}
\]
The circle criterion holds if \(\alpha > 0, \beta > 0, z \geq 0\) and \(0 < z \leq \alpha\).

The 45° Criterion states that class (1) is absolutely stable if the eigenvalues of the matrix pencil \((A - kbc^T)\), where \(k \geq 0\), lie inside a specific sector of the complex plane known as the 45° region. The 45° Criterion holds if \(\alpha > 0, \beta > 0, z \geq 0\) and \(0 < z \leq \alpha + \sqrt{\alpha^2 - 2\beta}\).

Theorem 1 improves upon the circle criterion and the 45° criterion. This is illustrated in the following examples.

#### 2.2.1. Example 1 (The eigenvalues of \(A\) are Real)

Consider class (1) with
\[
A = \begin{bmatrix} 0 & 1 \\ -2 & -3 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, c^T = \begin{bmatrix} z \\ 1 \end{bmatrix}
\]
Class (9) is absolutely stable according to Theorem 1, the 45° criterion and the circle criterion for the following ranges, see Table 1. Theorem 1 improves on both criteria.

#### 2.2.2. Example 2 (The eigenvalues of \(A\) are a Complex Conjugate Pair)

Consider class (1) with
\[
A = \begin{bmatrix} 0 & 1 \\ -5 & -4 \end{bmatrix}, b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, c^T = \begin{bmatrix} z \\ 1 \end{bmatrix}
\]
Class (10) is absolutely stable according to Theorem 1, the 45° criterion and the circle criterion for the following ranges, see Table 2. Theorem 1 improves on both criteria.

Piecewise quadratic common Liapunov functions for class (1) can be shown graphically by generating the level curves of the system, see Fig. 1. Here the Lur’e system of example 1 is examined. At first glance Fig. 1 appears to be a unic Liapunov function, however it is indeed a piecewise quadratic common Liapunov function. This may explain the difficulty in generalizing tractable stability conditions using unic Liapunov functions in higher dimensions. However the use of piecewise quadratic common Liapunov functions to establish stability for 2 dimensional systems shown in Lemmal can also be generalized for \(N\) dimensional systems.
### Table 1: Range of \( z \) for which class (9) is absolutely stable

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Range of ( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle Criterion</td>
<td>( 0 &lt; z \leq 3 )</td>
</tr>
<tr>
<td>45° Criterion</td>
<td>( 0 &lt; z \leq 5.236 )</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>( 0 &lt; z \leq 6 )</td>
</tr>
</tbody>
</table>

### Table 2: Range of \( z \) for which class (10) is absolutely stable

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Range of ( z )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circle Criterion</td>
<td>( 0 &lt; z \leq 4 )</td>
</tr>
<tr>
<td>45° Criterion</td>
<td>( 0 &lt; z \leq 6.449 )</td>
</tr>
<tr>
<td>Theorem 1</td>
<td>( 0 &lt; z \leq 8 )</td>
</tr>
</tbody>
</table>

### 3. Conclusion

In this paper a new stability criterion for 2-D nonlinear time varying Lur’e systems is presented. This new criterion improves on both the circle criterion and the 45° criterion and is illustrated using numerical examples. This result arose from determining the existence of a piecewise quadratic common Liapunov function for the system. It was also noted that piecewise quadratic common Liapunov functions of this form for 2-D systems appear graphically to be unci Liapunov functions while in fact they have a define quadratic nature.

### 3.1. Proof of Lemma 1

Let \( V(x) \) be of form (2) then

\[
x^T P_1 x = x^T P_2 x \quad \text{for all } x \quad \text{such that} \quad (b^T x)(c^T x) = 0
\]

**Lemma 2:**

\[
x^T \Delta x = 0 \quad \text{for all } x \quad \text{such that} \quad (b^T x)(c^T x) = 0
\]

where \( \Delta = \Delta^T \in \mathbb{R}^{n \times n} \) iff there exists a \( \delta \in \mathbb{R} \) such that

\[
\Delta = \delta (bc^T + cb^T)
\]

Therefore from (11) and Lemma 2

\[
(P_1 - P_2) = \delta (bc^T + cb^T) \quad \text{where} \quad \delta \in \mathbb{R}.
\]

Let \( P_0 = \frac{P_1 + P_2}{2} \) then \( P_0 = P_0^T \in \mathbb{R}^{2 \times 2} > 0 \) and by (14)

\[
P_1 = P_0 + \frac{\delta}{2} (bc^T + cb^T)
\]

\[
P_2 = P_0 - \frac{\delta}{2} (bc^T + cb^T)
\]

In order for \( V(x) \) to be a common Liapunov function for class (1), negative definite derivative conditions on \( V(x) \) require

\[
x^T P_1 b (c^T x) + x^T c b^T P_1 x \geq 0 \quad \text{for} \quad (b^T x)(c^T x) > 0
\]

Applying the S Procedure [16] to (17) and (18) yields

\[
P_1 b = \tau_1 b + \gamma_1 c
\]

\[
P_2 b = -\tau_2 b + \gamma_2 c
\]

for certain scalars \( \tau_1, \tau_2, \gamma_1, \gamma_2 \geq 0 \).

From (15) and (16)

\[
P_0 b = \left( \tau_1 - \frac{\delta}{2} \right) b + \left( \gamma_1 - \frac{\delta}{2} \right) c
\]

\[
P_0 b = \left( -\tau_2 + \frac{\delta}{2} \right) b + \left( \gamma_2 + \frac{\delta}{2} \right) c
\]

Since \( [b, c] \) are linearly independent this implies that \( \tau_1 + \tau_2 = \delta \) and \( \gamma_1 = \gamma_2 = \frac{\delta}{2} \). Hence

\[
0 \leq \tau_1 \leq \delta, \quad 0 \leq \delta \leq \gamma_1
\]

If \( \delta = 0 \), \( P_1 = P_2 \), \( V(x) \) is a common quadratic Liapunov function and the circle criterion holds, therefore we assume that \( \delta > 0 \). Since we can scale without loss of generality let \( \delta = 2 \).

Applying the S Procedure to (19) and (20)

\[
A^T P_1 + P_1 A + q_1 (bc^T + cb^T) \leq 0
\]

\[
A^T P_2 + P_2 A - q_2 (bc^T + cb^T) \leq 0
\]

where \( q_1, q_2 \geq 0 \).

From (15) and (16)

\[
A^T P_0 + P_0 A + A (bc^T + cb^T) + (bc^T + cb^T) A
\]

\[
+ \frac{q}{2} (bc^T + cb^T) + \frac{q^2}{2} (bc^T + cb^T) \leq 0
\]

\[
A^T P_0 + P_0 A - A (bc^T + cb^T) + (bc^T + cb^T) A
\]

\[
- \frac{q}{2} (bc^T + cb^T) + \frac{q^2}{2} (bc^T + cb^T) \leq 0
\]

where \( q = q_1 + q_2 \geq 0, \eta = \frac{q - q_1}{q_1 + q_2} \in [-1, 1] \)

Letting \( \tau_1 = 1 = \tau \) and \( \gamma_1 = 1 + \gamma \) (23) becomes

\[
P_0 b = \tau b + (\gamma + 1) c
\]

where \(-1 \leq \tau \leq 1 \) and \( \gamma \geq 0 \).
3.2. Proof of Theorem 1

Class (1) is absolutely stable if all the conditions of Lemma 1 are satisfied. Consider class (1). Let \( P_0 = \begin{bmatrix} p & z \\ z & 2 \end{bmatrix} \). Therefore from (5),

\[
P_0 b = \begin{bmatrix} z \\ 2 \end{bmatrix} = b + c
\]

From (10),

\[
A^TP_0 + P_0 A + N - M_q = \begin{bmatrix} 0 & p - qz \\ p - qz & -2q \end{bmatrix} \leq 0
\]

which clearly implies that \( p = qz \).

Also

\[
A^TP_0 + P_0 A + N + M_q = \begin{bmatrix} -4\beta z & qz - 2az - 4\beta \\ qz - 2az - 4\beta & -4(2\alpha - z) \end{bmatrix} \leq 0
\]

This holds if

\[
16\beta z(2\alpha - z) \geq (qz - 2az - 4\beta)^2
\]

This clearly implies that \( 2\alpha - z \geq 0 \).

Moreover \( P_0 \) is positive definite if and only if \( q > \sigma \).

Therefore,

\[
\sqrt{\beta z(2\alpha - z)} \geq \sigma - 4\alpha z - 8\beta
\]

which implies \( z^2 - 4\alpha z - 8\beta \leq 0 \).

If \( z^2 - 4\alpha z - 8\beta > 0 \) then

\[
(a) \ z > \frac{4\alpha + \sqrt{16\alpha^2 + 32\beta}}{2} \quad \text{or} \quad (b) \ z < \frac{4\alpha - \sqrt{16\alpha^2 + 32\beta}}{2}
\]

(a) implies that \( z > 4\alpha \) and (b) implies that \( z < 0 \) both of which are false since \( 0 < z \leq 2\alpha \).

Therefore there exists \( P_0 > 0 \) such that all the conditions of Lemma 1 are satisfied and class (1) is absolutely stable if \( 0 < z \leq 2\alpha \).

References


On the modeling of nonlinear multi-time-scale systems  
- A recursive order reduction procedure -  
Hajer Bouzaouache1,2, Nizar Rezgui1  
1 Laboratoire d'Etudes et de Commande Automatique des Processus (L.E.C.A.P.)  
Ecole Polytechnique de Tunisie, BP 748, 2078 La Marsa, Tunisia  
2 Institut Supérieur des Etudes Technologiques en Communication de Tunis  
hajer.bouzaouache@ept.rnu.tn

Abstract - Most industrial processes are inherently nonlinear and of high order, which make their implementation prohibitive from computational considerations, especially when it appears an interaction of dynamic variables in separate time-scales. That's why enormous effort has been spent on the development of techniques for systems order reduction. In this paper, an analytical and recursive order reduction approach for nonlinear multi-time-scale systems is presented. It offers an alternative route to simplify the studied systems and provides a reduced model described by polynomial state equations. This approach is based on the application of the singular perturbation method and the use of a powerful mathematical tool: the Kronecker product. Its useful notations and properties allow important algebraic manipulations and an interesting description not only for the studied systems but also for the derived reduced order slow subsystem model. The proposed recursive procedure has been tested on a numerical example and produced useful results.

Keywords – Large scale system, polynomial description, Kronecker product, time scale modeling, order reduction, recursive method, computational procedure.

1. Introduction

The mathematical description of the most industrial processes leads usually to a complex high order model. This complexity and increasing order of dynamic systems is due to the presence of some small parameters such as small time constants, masses, capacitances, etc.,[1,3]. Introducing the multi-time-scale property these systems simultaneously poses states with different speeds. Coupling of these states with each other makes the system analysis much more complex. But for analysis, controller synthesis and due to practical considerations, a reduced-order model is preferred. Indeed controller implementations usually rely on reduced order models to represent the behaviour of the process [2,4,6,8,14]. Fortunately, the singular perturbations theory, which have been studied extensively [1,2], gives us a powerful tool for order reduction and separation of time-scales studies. This singular perturbations approach, arising from an attempt to approximate a high-order system with another one of lower order, has proven to be a successful analytical tool that exploits directly the separation of system time scales, made explicit by the small singular parameter $\mu$. In this work, we propose a simplified and easy implemented approach for order reduction of higher order nonlinear continuous systems based on the description of these systems by an analytical control affine state space equation, using the Kronecker product [11].

This paper is organised as follows. In section 2, is presented an overview on the use of the singular perturbations theory in modeling nonlinear systems and formulating the problem statement. Te description of the studied systems is introduced in section 3. In section 4, the design recursive procedure for systems reduction is then proposed and derived accordingly the reduced order system which models the slow states of the studied system. Section 5, applies the developed procedure to a design example with four-time-scales. Finally some conclusions and comments are formulated.

2. Singular perturbation theory and problem statement

Due to not only theoretical interest, but also to the relevance of this topic to control engineering applications, the singular perturbations techniques have received a great deal of attention over the past two decades [1]. Several results are developed in the literature to cope with order reduction of linear systems [2,5] and even for particular classes of nonlinear systems [3,6,7] using the singular perturbations theory. Let us consider the general nonlinear singularly perturbed systems described by the following state equations:

\[
\begin{align*}
\dot{x} &= F(x,z,t), \quad x(t_0) = x_0, x \in \mathbb{R}^n \\
\mu \dot{z} &= G(x,z,t), \quad z(t_0) = z_0, z \in \mathbb{R}^n
\end{align*}
\]  

where the state variables are divided into $n_1$ slow states $x$ and $n_2$ fast states $z$ and $\mu$ is a positive small singular perturbed parameter such that $0 < \mu << 1$, which quantifies the speed ratio of the slow versus the fast dynamical phenomena of the process, $t \in \mathbb{R}^+$. One of the delicate problems is to explicit not defined at fortiori the parameter $\mu$ depending on the adopted representation of the studied system.

Let us introduce the following assumptions [10,12].

Assumption 1

The jacobian \( \frac{\partial G(x,z,t)}{\partial z} \) at \( \mu = 0 \) is non singular for all \( x, z \in D \), where \( D \) is a region of interest for the approximate analysis, [10].
Assumption 2
For all \( x, z \in D \), real parts of the eigenvectors of \( \frac{\partial \mathcal{G}(x,z,t)}{\partial z} \) at \( \mu=0 \) are smaller than a fixed negative number: \( \text{Re}\{\frac{\partial \mathcal{G}(x,z,t)}{\partial z}\} < c < 0 \)

When the Assumption 1 is satisfied, the slow reduced model is obtained by formally setting \( \mu=0 \) in equation (1) as:

\[
\begin{align*}
\mathbf{\tilde{X}} &= F(\mathbf{\tilde{X}}, \mathbf{\tilde{Z}}, t) \\
0 &= G(\mathbf{\tilde{X}}, \mathbf{\tilde{Z}}, t)
\end{align*}
\]

The reduced model will be determined by solving for \( \mathbf{\tilde{Z}} \) the equation (2-b). This appears to be easy in the case of linear systems but seems to be a difficult task in the general case of nonlinear systems. Several results are obtained when considering linear systems and particular class of nonlinear systems and some others in the case of nonlinear systems described by state equations with no direct transmission matrix. So the problem we are addressing in this paper can be stated as follows; given a high order multi-time scale system with nonlinear vector functions, how can we exploit the singular perturbation theory and the tensors representation to derive in a relatively simple way a simplified model and provide a computational procedure for this aim.

3. Description of the studied systems
We will focus on realistic models of \( q \) time-scales systems which involve interacting dynamic phenomena of widely different speeds and described by the following state separable form:

\[
\begin{align*}
\mathbf{\mu}_i \dot{X}_i &= F(X_i, \ldots, X_{i-1}, X_{i+1}, \ldots, X_q) + G(X_i, \ldots, X_{i-1}, X_{i+1}, \ldots, X_q) U \\
\vdots & \\
\mathbf{\mu}_q \dot{X}_q &= F(X_q, \ldots, X_{q-1}, X_1, \ldots, X_{q-2}) + G(X_q, \ldots, X_{q-1}, X_1, \ldots, X_{q-2}) U
\end{align*}
\]

where \( X_i \in \mathbb{R}^{n_i}, i = 1, \ldots, q \) are the state vectors, \( U \in \mathbb{R}^n \) is the control vector and \( F(\cdot), G(\cdot) \) \((i = 1, \ldots, q)\), the vector functions, continuously differentiable, are supposed analytic. Moreover, \( \mathbf{\mu}_i \), denotes small non-negative singular perturbation parameters such that \( \mu_i < \mu_k \) for \( k > i \), making the state variables to be ordered accordingly to their increasing speed. The system (4) can be studied like a multi-parameter system. \( \mu_i \) appears here for the aim of symmetry but it is going to be considered equal to 1. The simplification of the studied system (4) can be done gradually when posing successively \( \mu_i = 0, \mu_q = 0, \ldots \). It will appear a recurrence that makes possible to derive, going from the fastest to the slowest subsystem, the degenerate slow model. In each stage of reduction, we apply an analytical identification approach which is also introduced in this paper to calculate all the matrices appearing in the description of each reduced order model.

4. The proposed order-reduction procedure
The recursive scheme developed in this paper start with the following model obtained by rewriting the system (3) in a compact form:

\[
\begin{align*}
\dot{X}_n &= F_n(X_n, X_1) + G_n(X_n, X_1) U \\
\mu_n \dot{X}_1 &= H_n(X_n, X_1) + L_n(X_n, X_1) U
\end{align*}
\]

where \( X_n = [X_1^r \quad X_2^r \quad \cdots \quad X_q^r]^T \in \mathbb{R}^{q_n} \) is predominantly slow and \( X_q \) contains the \( n_q \) fastest state variables. \( n \) denotes the order of the full studied system defined by \( n = \sum_{i=1}^q n_i \).

\( F_n(\cdot), G_n(\cdot), H_n(\cdot) \) and \( L_n(\cdot) \), characterized from the functions \( F(\cdot)|_{\mu=0} \) and \( G(\cdot)|_{\mu=0} \)(see previous paper[15]), are also analytic functions and they admit a generalized Taylor series developments using the Kronecker product and power notations as follow:

\[
\begin{align*}
F_n(X_n, X_1) &= \sum_{i=1}^{n_i} F_{ni}(X_n, X_1[i+1-1]) \otimes X_1[j-1] \\
G_n(X_n, X_1) &= \sum_{i=1}^{n_i} G_{ni}(I_i \otimes X_1[i+1-1]) \otimes X_1[j-1] \\
H_n(X_n, X_1) &= \sum_{i=1}^{n_i} H_{ni}(X_n[i+1-1] \otimes X_1[j-1]) \\
L_n(X_n, X_1) &= \sum_{i=1}^{n_i} L_{ni}(I_i \otimes X_1[i+1-1] \otimes X_1[j-1])
\end{align*}
\]

where \( X_1[i] \) and \( X_1[j] \) are the \( i \)-th redundant Kronecker power of a \( n_i \) \((n_i = \sum_{i=1}^q n_i)\) dimensional vector respectively \( n_q \) dimensional vector. \( F_n, G_n, H_n, L_n \) and \( H_{ni} \) are constant matrices with appropriate dimensions. Generally, for modeling such studied systems, we have to truncate the previous development (5) at an arbitrary order \( r \). In practice, the studies [8,9,13,14] have demonstrated that the choice of the third order polynomial description \((r = 3)\) is general enough to model many physical processes. So, all the analytical functions (5) can be approximated in this case by vector polynomials.

First stage of reduction: Setting \( \mu_q = 0 \) and assuming that it exists a unique root solution of (4-b) in the interesting region which is considered here as a neighbourhood of the equilibrium taken to be the origin, the first degenerate subsystem verifying the assumption 1 and the assumption 2 is obtained by solving:

\[
\begin{align*}
\dot{\mathbf{\tilde{X}}}_n &= F_n(\mathbf{\tilde{X}}_n, \mathbf{\tilde{X}}_1) + G_n(\mathbf{\tilde{X}}_n, \mathbf{\tilde{X}}_1) U \\
0 &= H_n(\mathbf{\tilde{X}}_n, \mathbf{\tilde{X}}_1) + L_n(\mathbf{\tilde{X}}_n, \mathbf{\tilde{X}}_1) U
\end{align*}
\]

Then, after the first stage of system reduction the first obtained reduced model is described by the following polynomial state equation:

\[
\dot{\mathbf{\tilde{X}}}_n = \sum_{i=1}^{n_i} A_{ni} \mathbf{\tilde{X}}^{[i]} + \sum_{i=1}^{n_i} B_{ni}(I_n \otimes \mathbf{\tilde{X}}^{[i]}) U + O(\mathbf{\tilde{X}}^{[i]})
\]
where $\theta_{rs}^{(4)}$ is the fourth order error of (7).

The constant matrices $A_i$ and $B_i$ of appropriate dimensions are calculated when applying the identification approach presented below and summarize as follows:

- To get the property of full rank matrices, we first rewrite the equation (27) in the non-redundant form as:

$$
\hat{X}_s = \sum_{r=1}^{N} \hat{A}_r \hat{X}_s^{(r)} + \sum_{r=1}^{N} \hat{B}_r (I_s \otimes \hat{X}_s^{(r)}) U
$$

where $\hat{A}_s = A_s T$ and $\hat{B}_s = B_s (I_s \otimes T)$

- Considering the above notations and applying the vec function and the corresponding properties defined in [14], the equation (8) is rewritten as:

$$
vec(\hat{X}_s) = vec(\sum_{r=1}^{N} \hat{A}_r \hat{X}_s^{(r)} + \sum_{r=1}^{N} \hat{B}_r (I_s \otimes \hat{X}_s^{(r)})) U
$$

and rearranged as follows in order to extract the unknown matrices,

$$
vec(\hat{A}_s, \hat{X}_s) = (\hat{X}_s \otimes I_{n_s}) vec(\hat{A}_s)
$$

and

$$
vec(\hat{B}_s (I_s \otimes \hat{X}_s^{(r)})) = vec(\hat{B}_s) = (u'(I_s \otimes \hat{X}_s^{(r)}))' \otimes I_{n_s}
$$

where $$(16)$$

- Considering the above notations and applying the vec function, the proposition of the approach described by the polynomial state equations:

$$
\bar{X}_s = \sum_{r=1}^{N} \bar{A}_s \bar{X}_s^{(r)} + \sum_{r=1}^{N} \bar{B}_s (I_s \otimes \bar{X}_s^{(r)}) U
$$

and rearranged as follows to extract the unknown variables

$$
\bar{A}_s, \bar{X}_s
$$

and

$$
vec(\bar{B}_s (I_s \otimes \bar{X}_s^{(r)})) = vec(\bar{B}_s)
$$

where

$$
\bar{X}_s = [X_{s_1}, X_{s_2}, \ldots, X_{s_{n_s}}] \in \mathbb{R}^{n_s}
$$

and

$$
\bar{X}_s = \mathbb{R}^{n_s} \text{ with } n_s = \sum_{r=1}^{N} n_r
$$

After the calculus of the matrices $\hat{A}_s$ and $\hat{B}_s$ the next step is to extract again the fastest state variables given by (16) $\bar{X}_s$ and to derive the degenerate model that approximate the slowest state variables of the full studied systems where the original variables are finally approximated by:

$$
\bar{X}_s = [X_{s_1}, X_{s_2}, \ldots, X_{s_{n_s}}] \in \mathbb{R}^{n_s}
$$

For this aim, it is necessary to develop the terms $\hat{X}_s$ and $(I_s \otimes \hat{X}_s^{(r)})$ (19) in order to make to appear the slow state variables $\bar{X}_s$ and the fastest ones $\bar{X}_s$.

Consequently, the system (19) can be formulated by a similar description as (4) and (5).

It comes out that all the constant matrices $F_{s,k}, G_{s,k}, L_{s,k}$ and $H_{s,k}$ are then defined and we can derive the finally degenerate slow model:

$$
\bar{X}_s = \sum_{s_{s_1(s_k)}} A_{s_{s_1(s_k)}} \bar{X}_s^{(s_{s_1(s_k)})} + \sum_{s_{s_2(s_k)}} B_{s_{s_2(s_k)}} (I_s \otimes \bar{X}_s^{(s_{s_2(s_k)})}) U
$$

where $A_{s_{s_1(s_k)}} = \hat{A}_s$ and $B_{s_{s_2(s_k)}} = \hat{B}_s (I_s \otimes T)$

$\hat{A}_{s_{s_1(s_k)}}$ and $\hat{B}_{s_{s_2(s_k)}}$ are the solutions of the following equation:

$$
\bar{X}_s = \theta_{rs}^{(r)} \hat{X}_{s_{rs}}
$$

and

$$
\bar{X}_s = \theta_{rs}^{(r)} \hat{X}_{s_{rs}}
$$

So to derive the reduced order model, we have to determine the unknown vector $\theta_{rs}$ which needs the use of the Moore-Penrose pseudo-inverse $\theta_{rs}$ given by:

$$
\theta_{rs} = [\theta_{rs}, \theta_{rs}]^T
$$

The solution of the vectorial equation (17) is then equal to:

$$
\bar{X}_s = \theta_{rs}^{(r)} \hat{X}_{s_{rs}}
$$

Applying a random input to the system, the proposed approach provides the determination of the matrices $\hat{A}_s$ and $\hat{B}_s$, $i = 1, \ldots, r$ and so on the matrices $A_i$ and $B_i$ from (9).

Starting from the obtained reduced model (7) which contains $(q-1)$ state vectors of different speeds, the gradually application of the proposed approach of system reduction derives when going from the fastest to the slowest subsystem a reduced order slow subsystem model.

**Determination of the slow model after $(K+1)$ stages of reduction**

After this first stage of reduction, it appears from the above developments a recurrence. So, having assumption 1 and 2 verified in each stage, the repeated application of order reduction until the $K^{th}$ one leads to the following reduced model described by the polynomial state equations:

$$
\bar{X}_s = \sum_{s_{s_1(s_k)}} A_{s_{s_1(s_k)}} \bar{X}_s^{(s_{s_1(s_k)})} + \sum_{s_{s_2(s_k)}} B_{s_{s_2(s_k)}} (I_s \otimes \bar{X}_s^{(s_{s_2(s_k)})}) U
$$

where

$$
\bar{X}_s = [X_{s_1}, X_{s_2}, \ldots, X_{s_{n_s}}] \in \mathbb{R}^{n_s}
$$

and

$$
\bar{X}_s = \mathbb{R}^{n_s} \text{ with } n_s = \sum_{r=1}^{N} n_r
$$

After a random input to the system, the proposed approach provided the determination of the matrices $\hat{A}_s$ and $\hat{B}_s$, $i = 1, \ldots, r$ and so on the matrices $A_i$ and $B_i$ from (9).

Starting from the obtained reduced model (7) which contains $(q-1)$ state vectors of different speeds, the gradually application of the proposed approach of system reduction derives when going from the fastest to the slowest subsystem a reduced order slow subsystem model.
where computing all the factor scale $\mu_i$ are incorporated in the corresponding matrices. Good results are obtained when comparing the trajectories behavoutre of the initial system and the reduced obtained model (see the following illustrative example).

### 5. Numerical example

In order to demonstrate the efficiency of our proposed recursive procedure, we have considered the following nonlinear 7th order system with four time-scales.

\[
\begin{cases}
\dot{X}_{1,7} = F_{1,7}(X_{1,7}) + G_{1,7}(X_{1,7}, \dot{X}_{1,7}) U \\
0 = H_{1,7}(X_{1,7}, \dot{X}_{1,7}) + L_{1,7}(X_{1,7}, \dot{X}_{1,7}) U
\end{cases}
\]  

(23)

After a three stages of system reduction the implemented approach, when considering the input signal as defined in Fig. 1, allows the determination of the matrices $A_{ij}$ and $B_{ij}$ of the degenerate model. The obtained results show the effectiveness of the recursive order reduction procedure and the quality of the reduced model as illustrate in the figure 2.

with $\mu_1 = 1$; $\mu_2 = 0.3$; $\mu_3 = 0.1$; $\mu_4 = 0.03$

After a three stages of system reduction the implemented approach, when considering the input signal as defined in Fig. 1, allows the determination of the matrices $A_{ij}$ and $B_{ij}$ of the degenerate model. The obtained results show the effectiveness of the recursive order reduction procedure and the quality of the reduced model as illustrate in the figure 2.

with $\mu_1 = 1$; $\mu_2 = 0.3$; $\mu_3 = 0.1$; $\mu_4 = 0.03$

\[
A_{i} = \begin{bmatrix}
0.0509 & 1.5581 \\
1.5581 & 0.0936
\end{bmatrix}, \quad A_{ii} = \begin{bmatrix}
-0.1697 & -0.1161 & -0.1161 & 0.4522 \\
-0.0448 & 0.0005 & 0.0005 & 0.4064 \\
-0.0002 & -0.0002 & -0.0002 & -0.0002
\end{bmatrix}
\]

\[
B_{i} = \begin{bmatrix}
0.0988 & -0.0386 & -0.0386 & -0.1702 \\
-0.0386 & -0.1702 & -0.1702 & 0.2374 \\
0.0315 & 0.0012 & 0.0012 & -0.0214 & -0.0214 & -0.0214 & -0.2668 \\
0.0883 & 0.0451 & 0.0451 & 0.1746 \\
0.4828 & 0.0484 & 0.0484 & 0.8076 \\
0.4441 & -0.1968 & -0.1968 & -0.4601 & -0.1968 & -0.4601 & -0.4601 & -0.4601
\end{bmatrix}
\]

In this paper, an analytical recursive procedure has been proposed for order reduction of nonlinear multi-time-scale systems in which the nonlinearities are of polynomial type. It must be pointed out that for modeling a large class of nonlinear systems, the proposed approach has demonstrated its capabilities on identifying all the characteristic matrices of the determined nonlinear reduced order model. The procedure, computationally efficient, has been numerically tested on a nonlinear 7th order system and produced useful results. It appears to be a very powerful alternative to the general problem of modeling of complex systems and seems to correspond very well to industrial applications and power systems.

### References

[1] NAID D.S.

Singul\'ar perturbation and time scale in control theory and applications. An overview.


Singularity perturbations and order reduction in control theory. An overview. 6th IFAC World Congress, pp. 17-21, 1989.

Pittsburgh, 1975.


Multi-scale analysis of a power system.


[7] BENEFICE M. AND GASMI M.

On the use of an arrow form matrix for modeling and stability analysis of singularly perturbed nonlinear system.


[8] BOUZAOUACHE H. AND BENHADJ BRAIEK E.


[9] BOUZAOUACHE H. AND BENHADJ BRAIEK E.

Simplification du modèle non linéaire d’une machine synchronisée.


[10] TIRONOV A.N.


Kronerke product and matrix calculus in system theory.


Singularity perturbations and time scales in nonlinear models of power systems.


[14] BOUZAOUACHE H., BENHADJ BRAIEK E. AND BENEFICE M.

A reduced optimal control for nonlinear singularly perturbed systems.


[15] BOUZAOUACHE H.

Sur la représentation d’état des systèmes non linéaires singulièrement perturbés polynomial.

Conférence Tunisienne de génie électrique CTGE04, pp 420-423, 2004
Multi-objective Design Scheme for Robust Pole Assignment

Takanori Tagami† and Kenji Ikeda‡
†Dept. of General Education, Anan National College of Technology, Tokushima, Japan
‡Dept. of Information Science and Intelligent Systems, The University of Tokushima, Tokushima, Japan
Email: tagami@anan-nct.ac.jp, ikeda@is.tokushima-u.ac.jp

Abstract—In this paper, a design method of robust I-PD controller based on a pole assignment is proposed. The proposed design method is formulated as a two-objective optimization problem in which an issue of settling time and damping ratio is translated into a pole assignment problem. A genetic algorithm is applied to optimize this optimization problem because of its multiple search property. In order to investigate the effectiveness of the proposed design method, a magnetic levitation system is used as an example of an uncertain plant.

1. Introduction

PID-type controller [1] is effective enough to give desirable control performance for practical use. However, a design problem of PID-type control system is not always formulated as a concise convex problem. Moreover, taking into account the robustness for the uncertainty of actual plants, the design problem becomes more complex.

In practice, settling time and damping ratio are often important factors for the design of PID-type control systems. Since there is a trade-off between the settling time and the damping ratio, a controller is designed by trial and error via observing both properties. Thus, a design method based on the optimization of the settling time and the damping ratio is required to be established for practical use.

In this paper, we formulate the design problem of I-PD controller [2] as a two-objective optimization problem which optimize both the settling time and the damping ratio. Although the solution of the two-objective optimization problem is not a single one in general, we can obtain a set of solutions known as the Pareto-optimal solutions. If the Pareto-optimal solutions of this problem can be found efficiently, it will reduce the burden of controller designers. In order to formulate this design problem as a tractable optimization problem relevant to the settling time and the damping ratio. In the proposed design method, a genetic algorithm (GA) [3-5] is used for the optimization problem taking into account the advantages of its multiple search property. The uncertainties of the plant are represented as a polytope of polynomials, and a design cost is reduced by using the edge theorem [6].

In order to demonstrate the effectiveness of the proposed design method, a magnetic levitation system is used as an example of an uncertain plant.

2. Problem Statement

In this paper, a plant uncertainty is represented by using a polytope of polynomials. For the purpose of reducing a design cost, the edge theorem [6] is employed.

2.1. I-PD Control System

Suppose a plant can be modeled by an all-pole transfer function:

\[
g(s) = \frac{1}{a_0 + a_1 s + a_2 s^2 + \cdots + a_n s^n}
\]

where \( s \) is the Laplace operator.

Consider the control system with the I-PD controller [2] shown in Fig. 1.

\[
r(t) \quad e(t) \quad \frac{k_I}{s} \quad u(t) \quad g(s) \quad y(t)
\]

\[
k_P + k_D s
\]

Fig. 1 I-PD control system

From Fig. 1, the I-PD control system is described by

\[
e(t) = r(t) - y(t)
\]

\[
u(t) = \frac{k_I}{s}e(t) - (k_P + k_D s)y(t)
\]

\[
y(t) = g(s)u(t)
\]

where \( r(t) \) is the reference input, \( e(t) \) is the error signal, \( u(t) \) is the controller output, \( y(t) \) is the plant output. The variables \( k_I, k_P \) and \( k_D \) indicate the I-PD parameters. Hence, the transfer function of the closed-loop system is described as follows:

\[
H_{cl}(s) = \frac{g(s)k_I}{s + g(s)(k_I + k_P s + k_D s^2)}
\]

\[
= \frac{1}{1 + \frac{a_0 + k_P}{k_I} s + \frac{a_1 + k_D}{k_I} s^2 + \cdots + \frac{a_n}{k_I} s^{n+1}}
\]
By applying the I-PD control system to the plant that is modeled by an all-pole transfer function, we can also model the transfer function of the closed-loop system by means of an all-pole transfer function.

2.2. Representation of Model

Each trial of system identification will produce a distinct plant model. Thus, the plant can be represented by a set of models. The set of models is defined as a polytope of denominator polynomials as follows:

\[ A_i(s) = \lambda_1 A_1(s) + \lambda_2 A_2(s) + \cdots + \lambda_{N_0} A_{N_0}(s) \]  
\[ = a_{i0} + a_{i1} s + a_{i2} s^2 + \cdots + a_{in} s^n \]

where \( A_i(s) \), \( i = 1, \ldots, N_0 \) denotes a denominator polynomial of the plant model. The coefficient parameter \( \lambda_j \), \( \sum \lambda_j = 1, 0 \leq \lambda_j \leq 1, j = 1, \ldots, n \) represents the uncertainty. By using Eq. (8), the closed-loop system (6) is transformed by

\[ H_{y_2}(s) = \frac{1}{1 + \frac{\alpha_{i0} + k_p}{k_i} s + \frac{\alpha_{i1} + k_D}{k_i} s^2 + \cdots + \frac{\alpha_{in}}{k_i} s^{n+1}} \]  
\[ \theta = \text{arg min} \{ \text{Real}(s_i) \} \]  
\[ \lambda \leq \text{max} \{ \theta(s_i) \} \]  

The generating extreme points [6] of the plant denominator polynomials correspond with those of the polytope of characteristic polynomials.

2.3. Objective Functions

The purpose of the design problem is to find the I-PD parameters which minimize the following objective functions simultaneously:

\[ \text{minimize } \mathbf{F} = (F_1, F_2)^T \]  
\[ F_1 = \max_{1 \leq i \leq n} \{ \text{Real}(s_i) \} \]  
\[ F_2 = \max_{1 \leq i \leq n} \{ \theta(s_i) \} \]

where \( s_i \) indicates a pole of the closed-loop transfer function obtained from (9), \( \theta_i \) indicates an angle between the real axis and the straight line which connects the pole \( s_i \) with the origin. The objective function \( F_1 \) is the maximum value of the real part for all poles, which is concerned with the settling time. On the other hand, the objective function \( F_2 \) is the maximum value of the angle between the real axis and the straight line which connects the pole \( s_i \) with the origin, which is associated with the damping ratio.

Fig. 2 illustrates the objective functions \( F_1 \) and \( F_2 \) for a typical pole placement of an uncertain system.

3. Design Algorithm

The design algorithm of a robust I-PD controller based on the GA is summarized as follows:

**Fig. 2** The objective functions \( F_1 \) and \( F_2 \) for a typical pole placement of an uncertain system

**Step 1.** Generate a plant parameters set from system identifications. Then calculate the generators [6] and the discrete exposed edges from the identifying results.

**Step 2.** Let \( S_1 \) be a set of polytope of polynomials for all generators, let \( S_2 \) be a set of polytope of polynomials for all discrete exposed edges.

**Step 3.** Compute the I-PD parameters set which minimize the objective functions (11) and (12) for each element of the set \( S_1 \) by using GA explained following subsection.

**Step 4.** By using all I-PD parameters obtained from Step 3, check the objective function values for each element of the set \( S_2 \). If both objective function values are deteriorated by elements of the set \( S_2 \), the elements add to the set \( S_1 \). Then go to Step 3. Otherwise, go to Step 5.

**Step 5.** Select the desirable I-PD parameters by observing the actual experimental property.

In the proposed design method, the binary search method is employed as a selecting procedure of the desirable triplet of the I-PD parameters. A controller designer can easily obtain a desirable triplet of the I-PD parameters in only a few trials.

3.1. Procedure of GA

The continuous generation model with a Pareto-optimal preservation strategy is employed. In this model, Pareto-optimal solutions are always forced to appear in the following generation. The proposed procedure of GA consists of the following steps.

**Step 1.** Set a generation number \( \tau = 0 \). Randomly generate an initial population \( P(0) \) of \( M \) individuals.

**Step 2.** Calculate the fitness of each individual in the current population \( P(\tau) \) according to the multi-objective ranking method [3].
Step 3. Generate a new population \( P'(\tau) \) as follows:
select individuals from \( P(\tau) \);
recombine them using a crossover operator.

Step 4. Apply a mutation operator to the individuals ac-
cording to the mutation rate.

Step 5. Calculate the fitness both \( P(\tau) \) and \( P'(\tau) \). Select
the top \( M \) individuals from all population members on
the basis of their fitness.

Step 6. If a terminal condition is satisfied, stop and return
the superior individual. Otherwise, set \( \tau = \tau + 1 \) and
go to Step 2.

In this procedure, the current population size is always
constant \( M \). Here, to avoid the rapid loss of population di-
versity, multiple equivalent individuals are eliminated from
the current population.

4. Application to Magnetic Levitation System

In order to investigate the effectiveness of the proposed
design method, we applied the proposed design method to
a magnetic levitation system.

4.1. Magnetic Levitation System

The experimental system is shown in Fig. 3. The posi-
tion of the steel ball is measured by a laser gap sensor, and
is sent to the computer through the A/D converter. The sig-
nal from the computer is sent to the voltage control circuit
through the D/A converter, and it becomes a voltage of the
coil edge. The I-PD controller is implemented as a program
on the computer.

The control objective is to maintain the position of the
steel ball to the specified position. Since the magnetic lev-
itation system is modeled by a 3rd order nonlinear system,
the change of the equilibrium point causes the uncertainty
of the plant. Moreover, heat fluctuations and hysteresis
of the magnetic materials also cause the plant uncertainty.

4.2. Plant Model

Suppose the plant can be modeled as an all-pole 3rd or-
der transfer function. The transfer function of the plant is
expressed as follows:

\[
g(s) = \frac{1}{a_0 + a_1 s + a_2 s^2 + a_3 s^3}
\]  (13)

The generating extreme points of the polytope of poly-
nomials are obtained from several system identifications.

4.3. Experimental Results

In the design of the I-PD controller, each I-PD parameter
\( k_I, k_P \) and \( k_D \) is subject to interval constraint. We assume
that the constraints of the I-PD parameters are given as follows:

\[
\begin{align*}
0.0 & \leq k_I & \leq 10000.0 \\
0.0 & \leq k_P & \leq 10000.0 \\
0.0 & \leq k_D & \leq 10000.0
\end{align*}
\]  (14)

The following GA parameter specifications are em-
ployed for all experiments.

<table>
<thead>
<tr>
<th>Population size:</th>
<th>96</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mutation rate:</td>
<td>0.1</td>
</tr>
<tr>
<td>Terminal condition:</td>
<td>1000</td>
</tr>
</tbody>
</table>

Fig. 4 shows the result by using the proposed design
method. In the figure, the axis of abscissa indicates the
value of the objective function \( F_1 \), the axis of ordinate in-
dicates the value of the objective function \( F_2 \). The dots
indicate the solutions obtained from the final population.
From the results of preliminary experiments, we experi-
mented again in the search region \( R_1 \). Fig. 5 shows the
result in the search region \( R_1 \). The solutions are sorted into
the increasing ordered value of the objective function \( F_1 \).
It is clear that the proposed design method can seek for the
widely distributed solutions on the trade-off surface.

Table 1 shows the I-PD parameter values of the repre-
sentative points shown in Fig. 5.

<table>
<thead>
<tr>
<th>Values of the designed I-PD parameters</th>
<th>( k_I )</th>
<th>( k_P )</th>
<th>( k_D )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_{48} )</td>
<td>13.2341</td>
<td>4.6395</td>
<td>0.2876</td>
</tr>
<tr>
<td>( p_{60} )</td>
<td>10.2407</td>
<td>3.8779</td>
<td>0.2347</td>
</tr>
<tr>
<td>( p_{72} )</td>
<td>8.3628</td>
<td>3.4276</td>
<td>0.2032</td>
</tr>
<tr>
<td>( p_{96} )</td>
<td>6.3405</td>
<td>3.1438</td>
<td>0.1925</td>
</tr>
</tbody>
</table>

Fig. 6 shows the actual step response by applying the
triplet of the I-PD parameters \( p_{60} \) to the experiment system.
Fig. 7 shows the pole placements by applying the triplet of
the I-PD parameters \( p_{60} \). By using the binary search
method, the final I-PD parameters are determined from the
results of observing the settling time and the damping ratio.
In this case, the I-PD parameters \( p_{60} \) can be regarded as the
desirable solution.
5. Conclusion

In this paper, we have proposed the design method of I-PD controller for an uncertain plant. The proposed design method was defined as a two-objective optimization problem based on both of the settling time and the damping ratio. The uncertainties of the plant were represented as a polytope of polynomials, and a design cost was reduced by using the edge theorem. The genetic algorithm has been applied to optimize this problem because of its multiple search property. In the experiments, a magnetic levitation system has been used as an example of an uncertain plant. When the proposed design method is used in practical applications, the controller designer can easily find a desirable solution from among the Pareto-optimal set. Hence, the efficient design of an I-PD control system can be obtained by using the proposed design method.

Acknowledgments

This research was supported by the Ministry of Education, Culture, Sports, Science and Technology, Grant-in-Aid for Scientific Research, (KAKENHI 17700249).

References

Design of Formal Linearization and Observer for Time-Delay Nonlinear Systems

Kazuo Komatsu† and Hitoshi Takata‡

†Department of Information and Computer Sciences, Kumamoto National College of Technology
Suya 2659-2, Koshi, Kumamoto 861-1102 JAPAN
‡Department of Electrical and Electronics Engineering, Kagoshima University
Korimoto, Kagoshima 890-0065 JAPAN
Email: komatsu@cs.knct.ac.jp, takata@eee.kagoshima-u.ac.jp

Abstract—This paper is concerned with a design of linearization for multidimensional nonlinear systems with delays in state variables. A systematic design of the linearization is presented and its inversion is easily obtained by the solution of the linearized systems. As one of the applications of the linearization, a time-delay nonlinear observer is also designed. Numerical experiments show that the proposed linearization and observer are effective for time-delay nonlinear systems.

1. Introduction

We encounter nonlinear systems with delays in state variables in various fields and their treatments are usually difficult for their nonlinearities. To apply the linear systems theories for nonlinear time-delay systems, it is necessary to use linearization and some studies have been already reported [1, 2, 3, 4].

In this paper we present a design of a linearization for multidimensional nonlinear systems with delays in state variables and its application to an observer. Introducing a polynomial linearization function and a time-delay operator, time-delay nonlinear systems are transformed into time-delay linear ones with respect to the linearization function by using Taylor expansion. Its inversion is simple by extracting some elements of the linearization function. The advantage of this approach is easy and systematic to linearize time-delay nonlinear systems. As an application of this linearization technique, nonlinear observer is synthesized.

Numerical simulations show that these linearization and observer are effective for time-delay polynomial type nonlinear systems.

2. Formal Linearization

Assume that a time-delay nonlinear system is described by

\[ \Sigma_1 : x(t) = f(x(t)) + g(x(t-\ell)) \]  

where \( t > 0 \) denotes time, overdot represents derivative with respect to \( t \), \( \ell \in (0, \infty) \) is the system delay, \( x = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n \) is a state variable, \( f = [f_1, f_2, \ldots, f_n]^T \in \mathbb{R}^n \) and \( g = [g_1, g_2, \ldots, g_n]^T \in \mathbb{R}^n \) are sufficiently smooth nonlinear functions. The system initial condition is given by

\[ x(\theta) = \varphi(\theta) (-\ell \leq \theta < 0), \]

\[ x(0) = x_0. \]

In this paper we exploit a formal linearization of polynomial type using Taylor expansion up to the N-th order [5, 6, 7]. We define an N-th order linearization function \( \phi(\cdot) = \phi(x(\cdot)) \) which consists of polynomials by

\[ \phi = [\phi_1, \phi_2, \ldots, \phi_n, \phi_n(N^{th} - 1)]^T \]

\[ = [T_{(10-0)}(x), T_{(01-0)}(x), \ldots, T_{(00-1)}(x), T_{(11-0)}(x), T_{(10-1)}(x), \ldots, T_{(11-N)}(x), \ldots, T_{(N-1)}(x)]^T \]

where

\[ T_{(i_1-\ell)}(x) = \prod_{j=1}^{n} x_{i_j}^{\ell}. \]

Let a time-delay linear operator \( \delta_t \) be

\[ \delta_t[x_i(t)] = x_i(t-\ell) \quad (i = 1, \ldots, n) \]

which delays the time \( \ell \) from \( x_i(t) \). In vector expression, it is represented as

\[ \delta_t[x(t)] = x(t-\ell). \]

To these linearization function and time-delay operator, we derive the derivative of each element of \( \phi \) along with solution of the given nonlinear system(Eq. (1)) as

\[ \phi_{\alpha}(x) = T_{(r_1-\ell)}(x) \]

\[ = \left( \frac{d}{dt} x_1^{r_1} \right) x_2 \cdots x_{n-1} x_n + \cdots + x_1^{r_1} x_2^{r_2} \cdots x_{n-1}^{r_{n-1}} \frac{d}{dt} x_n^{r_n} \]

\[ = \sum_{k=1}^{n} r_k f_k(x(t)) + g_k(x(t-\ell)) \frac{T_{(r_1-\ell)}(x)}{x_k}, \quad (3) \]

\[ \alpha = \alpha(r_1, \cdots, r_n). \]
Note that Taylor expansion up to the N-th order derives
\[
f_k(x(t)) = \left[p_k 1, p_k 2, \cdots, p_k (N+1) n \right] \phi(x) + \text{higher order}
\]
where
\[
p_k j = \frac{\partial^{r_1+r_2+\cdots+r_n}}{\partial x_1^{r_1} \partial x_2^{r_2} \cdots \partial x_n^{r_n}} f_k(x) \bigg|_{x=0},
\]
and
\[
g_k(x(t-\ell)) = g_k(\delta x(t)) = \left[q_k 1(\delta), q_k 2(\delta), \cdots, q_k (N+1) r(\delta) \right] \phi(x) + \text{higher order}
\]
where
\[
q_k j(\delta) = \frac{\partial^{r_1+r_2+\cdots+r_n}}{\partial x_1^{r_1} \partial x_2^{r_2} \cdots \partial x_n^{r_n}} g_k(\delta x) \bigg|_{x=0}.
\]

From Eqs.(3), (4) and (5), it follows that
\[
\dot{\phi}\alpha(x) = r_1 [p_{11} + q_{11}(\delta), \ p_{12} + q_{12}(\delta), \cdots, \ p_{1(N+1)-1} + q_{1(N+1)-1}(\delta)] \phi(x) x_1^{r_1-1} x_2^{r_2} \cdots x_n^{r_n} + r_2 [p_{21} + q_{21}(\delta), \ p_{22} + q_{22}(\delta), \cdots, \ p_{2(N+1)-1} + q_{2(N+1)-1}(\delta)] \phi(x) x_1^{r_1} x_2^{r_2-1} \cdots x_n^{r_n} + \cdots + r_n [p_{n1} + q_{n1}(\delta), \ p_{n2} + q_{n2}(\delta), \cdots, \ p_{n(N+1)-1} + q_{n(N+1)-1}(\delta)] \phi(x) x_1^{r_1} x_2^{r_2} \cdots x_n^{r_n-1} + \text{higher order}
\]
\[
= \left[G_{\alpha 1}(\delta), \cdots, G_{\alpha \beta}(\delta), \cdots, G_{\alpha (N+1)-1}(\delta) \right] \phi(x) + \epsilon_\alpha(\delta, x)
\]
(6)
where
\[
G_{\alpha \beta}(\delta) = \sum_{k=1}^{n} r_k \left[p_k \beta(s_1-1,s_2-1,\cdots,s_n-1) \right] + q_k \beta(s_1-1,s_2-1,\cdots,s_n-1,1,\cdots,1)(\delta) \right],
\]
\[
\beta = \beta(s_1, \cdots, s_n),
\]
\[
p_k \beta + q_k \beta = \begin{cases} p_k j + q_k j & (\beta = j) \\ 0 & (\beta \neq j) \end{cases}
\]
The \(\epsilon_\alpha(\delta, x)\) is the error term whose order includes higher than \(N\) with respect to some \(x_i\) \((i = 1, \cdots, n)\). Using the above functions (Eqs. (4) and (5)), \(\phi\) (Eq. (3)) is approximated by
\[
\phi(x) = A(\delta)\phi(x)
\]
(8)
where
\[
A(\delta) = [G_{\alpha j}(\delta)] \in R^{[N(N+1)-1] \times [N(N+1)-1]}.
\]

Thus a formal time-delay linear system is derived by
\[
\dot{z}(x) = A(\delta) z(x),
\]
(9)
\[
z(x(\theta)) = \phi(\varphi(\theta)) \quad (-\ell \leq \theta < 0),
\]
\[
z(0) = \phi(x_0).
\]
Its inversion is simply obtained as follows. From Eq.(2), an approximated value \(\hat{x}(t)\) is
\[
\hat{x}(t) = [I \ 0 \ \cdots \ 0] \phi(x(t)) = [I \ 0 \ \cdots \ 0] z(t)
\]
(10)
where \(I\) is an \(n \times n\) unit matrix.

3. Time-Delay Nonlinear Observer

As an application of this method, we synthesize a time-delay nonlinear observer. The system is the same as Eq. (1):
\[
\Sigma_1 : \dot{x}(t) = f(x(t)) + g(x(t-\ell))
\]
(11)
and a measurement equation is assumed to be
\[
y(t) = H(\delta) \phi(t) \in R^m
\]
(12)
where \(H(\delta) \in R^{m \times m}\) could be a matrix function with respect to the time-delay operator \(\delta\).

The time-delay nonlinear system (Eq. (11)) is transformed into the time-delay linear one (Eq. (9)) by the formal linearization as mentioned above. To these time-delay linear systems (Eqs. (9) and (12)), we can apply the linear observer theory [8], so that an observer is obtained as
\[
\dot{\hat{x}}(t) = A(\delta) \hat{x}(t) + L(y - \dot{\hat{y}}),
\]
(13)
\[
\dot{\hat{y}} = H(\delta) \hat{z}(t)
\]
where \(L \in R^{n \times m}\) is the observer gain.

This observer asymptotically converges if we could set the gain \(L\) such that all poles of \(A(\delta) - LH(\delta)\) have negative real parts for any \(\delta\). From the inversion of Eq.(10), the estimate \(\hat{z}(t)\) is obtained by
\[
\hat{z}(t) = [I \ 0 \ \cdots \ 0] \hat{x}(t).
\]

4. Numerical examples

We illustrate numerical experiments of the formal linearization and the time-delay nonlinear observer.

4.1. Formal Linearization

We consider the following multidimensional time-delay nonlinear system:
\[
\begin{cases} 
\dot{x}_1(t) = -2x_2(t) \\
\dot{x}_2(t) = -x_2(t) - x_1^2(t-\ell),
\end{cases}
\]
(14)
\[
\varphi(\theta) = x_0 \ (-\ell \leq \theta < 0),
\]
When the order of the linearization function is \( N = 2 \), the linearization function (Eq. (2)) is
\[
\phi = [x_1, x_2, x_1 x_2, x_1^2, x_1 x_3, x_1 x_2 x_3, x_1^2 x_2, x_1^3, x_1 x_2^2, x_1^2 x_2^2, x_1^3 x_2]^T
\]
and \( \phi_\alpha \) (\( \alpha = 1, \ldots, 8 \)) of Eq. (6) are
\[
\phi_1 = \frac{d}{dt} x_1(t) = \dot{x}_1 = -2 x_2(t) = -2 \phi_2,
\]
\[
\phi_2 = \frac{d}{dt} x_2(t) = \dot{x}_2(t) = -x_2(t) - \delta_1^2 x_1^2(t) = -\phi_2 - \delta_1^2 \phi_4,
\]
\[
\phi_3 = \frac{d}{dt} x_1(t) x_2(t) = \dot{x}_1(t) x_2(t) + x_1(t) \dot{x}_2(t)
= -2 x_2^2(t) - x_1(t)(x_2(t) + \delta_1^2 x_1^2(t)),
\]
Thus the formal linear system (Eq. (8)) is
\[
\begin{pmatrix}
0 & -2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & -\delta_1^2 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & -2 & 0 & 0 & 0 \\
0 & 0 & -4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & -2 & 0 & -2 \delta_1^2 & 0 & 0 \\
0 & 0 & 0 & 0 & -2 & 0 & -1 & 0 \\
0 & 0 & 0 & 0 & 0 & -4 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -2 & 0
\end{pmatrix} \phi =
\begin{pmatrix}
0 \\
-2 \phi_2(t) \\
-\phi_3(t) - 2 \phi_2(t) \\
-4 \phi_3(t) \\
-2 \phi_3(t) - 2 \phi_2(t) \phi_4(t) - \ell \\
-2 \phi_3(t) - 2 \phi_2(t) \phi_4(t) - \ell \\
-2 \phi_4(t) - \phi_7(t) \\
-2 \phi_2(t)
\end{pmatrix}
\]
(15)
and the initial condition is
\[
\phi(\theta) = [x_1(0), x_2(0), x_1(0)x_2(0), x_1^2(0), x_2^2(0), x_1(0)x_2^2(0),
\]
\[
x_2^2(0), x_1(0)x_2(0), x_2(0), x_1(0)x_2(0), x_2(0), x_1(0)x_2(0)]^T (-\ell \leq \theta \leq 0)
\]
in this problem. Solving a formal linear system (Eq. (15)), the approximated solution \(\hat{x}(t)\) of the nonlinear system (Eq. (14)) is obtained by the inversion (Eq. (10)).

Figs. 1 and 2 show the true values \(x(t)\) which are solutions of Eq.(14), and the approximated values \(\hat{x}(t)\) \((i = 1, 2)\) when the order of the linearization function is varied as \(N = 1\) to 3. Fig. 3 depicts these errors by
\[
J(t) = \int_0^t \left((x_1(\tau) - \hat{x}_1(\tau))^2 + (x_2(\tau) - \hat{x}_2(\tau))^2\right) d\tau.
\]

### 4.2. Time-Delay Nonlinear Observer

We consider a following scalar system:
\[
\dot{x}(t) = -x(t) - 2x(t - \ell) - x^2(t) - x^2(t - \ell), \tag{16}
\]
\[
\phi(\theta) = x_0 \ (-\ell \leq \theta < 0),
x_0 = 0.5, \ \ell = 0.2
\]
and a measurement equation:
\[
y(t) = H(\delta)\phi(t),
\]
\[
H(\delta) = \begin{pmatrix}
-1 + \delta & \frac{1}{2}\delta^2 & 0 \\
0 & -1 + \delta & \frac{1}{2}\delta^2 \\
0 & 0 & -1 + \delta
\end{pmatrix}
\]

We set the parameters for the time-delay nonlinear observer as \(N = 3\) by
\[
L = \begin{pmatrix}
-2 & 0 & 0 \\
0 & -4 & 0 \\
0 & 0 & -6
\end{pmatrix}
\tag{17}
\]

In this case, the poles of
\[
A(\delta) - LH(\delta) = \begin{pmatrix}
-3 & -1 & 0 \\
0 & -6 & -2 \\
0 & 0 & -9
\end{pmatrix}
\]
are \((-3, -6, -9)\). Thus, an estimate of this observer (Eq. (13)) asymptotically converges to the true value.

Fig. 4 shows the true value \(x(t)\) and the estimate \(\hat{x}(t)\) when the unknown initial value is \(\hat{x}(0) = 0.8\).

### 5. Conclusions

We have considered a design of formal linearization method for multidimensional time-delay nonlinear systems using Taylor expansion. This method has a systematic procedure to linearize a time-delay nonlinear systems and enables us to be easy to apply linear system theories. As an application of this method, we synthesize a time-delay nonlinear observer. Numerical experiments show that accuracy of this method is improved as the order of the linearization function increases.

---

[References]


Chaos in dynamic atomic force microscopy

Arvind Raman† and Shuiqing Hu†
† School of Mechanical Engineering, Purdue University
West Lafayette, IN 47907, USA
Email: raman@ecn.purdue.edu, shuiqing@purdue.edu

Abstract—Deterministic chaos underpins the dynamics of many nonlinear systems that display a very high degree of sensitivity to initial conditions. In this article we summarize our recent experimental results on chaos in Tapping Mode Atomic Force Microscopy (TM-AFM) [1] and discuss the theoretical implications of the experimental findings. In particular we find that for softer cantilevers, the sudden onset of subharmonics or chaos accompanies the jump of the nanoscale tip from the attractive to the repulsive regime of interaction potential. Chaotic oscillations also occur under conditions of hard tapping. The results cast new light on the mechanisms through which the energy of the harmonically excited AFM microcantilever “scatters” or leaks into subharmonic sidebands and continuous chaotic spectra under the action of the nonlinear short-range tip-sample interaction forces.

1. Introduction
The observation of chaos weather patterns, speech patterns, stock market, and quantum physics has had a profound effect on these fields. In a recently published work [1], the authors observed and characterized chaos in the oscillations of nanoscale tips in Amplitude Modulated Atomic Force Microscopy (AM-AFM). These results suggest that “deterministic” uncertainty (chaos) may come into play in nanoscale metrology using AFM in addition to uncertainties due to statistical thermodynamics and quantum mechanics. Furthermore, it was shown that the presence of chaos can lead to errors while imaging samples using dynamic AFM, thereby introducing an element of deterministic uncertainty in nanometrology. In this article, we summarize some of these recent experimental results and discuss the theoretical implications of the recent experimental findings.

Prior to the present work, both Burnham et al [2] and Salapaka et al [3] have reported complicated microcantilever dynamics in experiments where vibrating samples were made to impact a stationary microcantilever. The present work deals more directly with nanoscale imaging and spectroscopy in dynamic AFM where the microcantilever is driven near its fundamental resonance.

Several theoretical models of AM-AFM are capable of predicting the onset of chaos.

1. When the cantilever is brought very close to the sample say within a few nanometers, the tip oscillates in a twin-well potential arising from the tip-sample interaction potential and the elastic potential of the cantilever. In such a situation it is well known that a periodic excitation can lead to homoclinic chaos [4] [5].

2. When the cantilever is further away from the sample (tens of nanometers), the cantilever oscillates in a single well potential. This is the most likely situation in realistic AFM experiments. Vibro-impact models with coefficient of restitution have been proposed to model this behavior [3] [6] [7]. In these models the stability of grazing trajectories is examined through the global dynamics of a Poincaré return map. Several bifurcations and chaotic attractors may be found in such models depending on the parameter values.

3. If one considers physics based tip-sample interactions such as the Derjaguin-Müller-Toporov (DMT) or Johnson-Kendall-Roberts (JKR), these models are either non-smooth or discontinuous. The dynamics of non-smooth oscillations is known to be complex, featuring grazing bifurcations and chaos [8]. Such non-smooth interactions may be able to capture the extremely high tip-sample interaction force gradients encountered in experiments.

4. Even with a smooth interaction potential such as the Lennard-Jones potential, it is possible that chaotic dynamics arise at sufficiently high excitation amplitudes for single-degree of freedom models [9].

5. Finally given that the cantilever is a continuous dynamic system, several different modes can participate in the response. The nonlinear coupling of these different modes may also lead to complex dynamics [10].

In this paper, we do not discuss homoclinic chaotic dynamics in the twin-well potential regime because this is a relatively uncommon situation in AM-AFM under ambient conditions where the cantilever is kept at least 10 nm away from the sample, typically 25-50nm.

2. Experimental Setup and Procedure
The experimental setup consists of a commercial AFM system, together with an external lock-in amplifier, and a high-speed data acquisition system. To minimize the effects of water meniscus, the experiments are performed on the hydrophobic surface of freshly cleaved HOPG (highly ordered pyrolytic graphic) and the environmental chamber is flushed with dry nitrogen. The driving frequency of the cantilever is controlled by the Signal Recovery™ lock-in amplifier via a general-purpose interface bus (GPIB) controller. To acquire simultaneously the real time data of the input piezo excitation and tip
deflection signal, two, single-channel National Instruments (NI) 5911 data acquisition boards are used. The sampling frequency is 12.5 MHz at 11 bits resolution.

In what follows, we will focus our attention mainly on experiments performed under dry nitrogen using the set of Nanosensor Force Modulation silicon microcantilevers (ω0=66.3 kHz) when driving frequency Ω = ω0 and with a setpoint amplitude of 80 nm. The phenomena observed in this set of microcantilevers under these operating conditions appear to span the phenomena observed for all other sets of microcantilevers. Details of experimental results of other cases can be found in our recent paper [1].

The goal of the experimental procedure is to demonstrate the existence of chaos and its effects on nanometry within the context of AM-AFM. Accordingly, for each microcantilever, the tip amplitude at the driving frequency (setpoint amplitude) is maintained constant via the lock-in amplifier and the AFM control system. The driving amplitude of the dither piezo is gradually increased in increments of 0.49 nm by increasing the voltage supplied to the dither piezo; this effectively decreases the setpoint amplitude ratio. At each driving voltage, both the tip oscillation and piezo excitation signal are recorded, and if needed the sample surface is scanned. These experiments are performed for two setpoint amplitudes, namely 80 nm, and 140 nm for each of the thirty or so microcantilevers tested. Moreover these experiments are performed for cases when the driving frequency (Ω) is just below, exactly at, and just above the fundamental resonance frequency (ω0) of each microcantilever, or for Ω=0.98 ω0, Ω=ω0, and Ω=1.02 ω0. Finally each experiment was repeated six times to verify the repeatability of the results. The observed phenomena are very robust over the six repeats of each experiment, and also consistent between the approximately ten microcantilevers tested within each set. This represents a wide range of microcantilever bending stiffnesses and resonance frequencies. Except for very stiff, and high frequency microcantilevers (such as the set of TESP microcantilevers, resonance frequency ω0=310 kHz), all other microcantilevers tested undergo chaotic oscillations upon transition from the non-contact to tapping regimes for certain operating parameters.

3. Experimental Results

In order to detect rigorously the presence of chaos in the tip oscillation time series data at each driving amplitude, we use two different time series analysis methods to detect the presence of chaos. A hallmark of chaotic dynamics is its sensitive dependence on initial conditions, as indicated by the presence of a positive Lyapunov exponent—a measure of the exponential divergence of nearby states. Positive Lyapunov exponent is indicative of chaos. The “lyp_k” routine of TISEAN 2.1 Nonlinear Time Series Analysis package is used to estimate the largest Lyapunov exponent of the time series vibration data at each driving amplitude. Titration of chaos with added noise [11] can robustly detect chaos with short noisy experimental data. Specifically this method can detect the nonlinearity in the time series by comparing short time prediction results of nonlinear and linear Volterra-Wiener-Korenberg series. White noise of increasing standard deviation is added to the time series until its nonlinearity cannot be detected by a particular numerical indicator. We denote this value of noise limit (NL) as the “noise ceiling” of the signals corresponding to the amount of chaos. NL > 0 indicates chaos; NL = 0 indicates that the data series either is not chaotic or the chaotic component is already neutralized by the background noise in the original time series. The results shown in Fig. 1 clearly show that the largest Lyapunov exponent and noise limit (NL) becomes positive at 37.2% setpoint amplitude ratio after the tip first transitions from the non-contact to the tapping regime. Following this transition, the tip continues to oscillate chaotically for driving amplitudes between 1.23nm and 2.70nm, then suddenly transitions to a period-2 oscillation and becomes chaotic again when the driving amplitude is larger than 4.66nm. Clearly both the titration of chaos method and the Lyapunov exponent method yield the same results and confirm that chaotic oscillations set in at the transition from non-contact to tapping regime of oscillation (at 37.2% amplitude setpoint ratio) and then again under hard tapping conditions.

FIG. 2. Tip deflection phase portrait (left) and power spectral density (right) versus dither piezo drive amplitude of (a) 0.49nm (period-1, noncontact regime), (b) 1.97nm (chaotic, tapping regime) and (c) 14.76nm (chaotic, tapping regime). The vertical

FIG. 1. Largest Lyapunov exponent and noise limit versus dither piezo driving amplitude for one of the Force Modulation (ω0=66.3 kHz) microcantilevers in a dry nitrogen environment over freshly cleaved HOPG.
line in the phase portrait represents the position of sample surface. $\Omega$ denotes driving frequency, $2B$ denotes the second bending mode frequency of the microcantilever. In phase portrait, velocity is computed by differentiating the tip vibration signal using moving average, and the hatched square represents the position of sample surface.

![FIG. 3. 3D Topology image with driving amplitudes of (a) 0.49nm (period-1, noncontact regime), (b) 1.97nm (chaotic, tapping regime) and (c) 14.76nm (chaotic, tapping regime).](image)

We will now explore the consequences of imaging the freshly cleaved atomically flat HOPG surface in different regimes. For the driving amplitude 0.49 nm the tip is in a period-1 oscillation state, in the non-contact regime. The tip vibration power spectral density shows superharmonics, but the tip motion is still period 1 as shown in the phase portrait plot in Fig. 2(a). Fig. 3(a) shows the 3D topology map of the freshly cleaved HOPG surface acquired under these conditions. Imaging the sample in this regime produces accurate topology images of freshly cleaved HOPG. Fig. 2(b) shows tip deflection phase portrait and power spectral density at the drive amplitude of 1.97 nm when the tip is in its first chaotic oscillation state at the transition from non-contact to tapping regime. The spectrum now contains superharmonics, subharmonics, side bands and the noise floor increases as seen in this power spectrum. Note that a significant sixth harmonic of the driving frequency also appears in tapping regime near the second bending mode resonance frequency (denoted by ‘$2B$’ in Fig. 2(b)). This is attributed to a nonlinear coupling between two different modes of the continuous microcantilever beam - the fundamental mode and the second bending mode whose resonance frequency is close to six times that of the fundamental mode. This condition is also referred to as a 1:6 internal resonance [10]. The topology map at driving amplitude of 1.97 nm shown in Fig. 3(b) is noisier due to chaotic nature of the tip oscillation. Finally, when the driving amplitude is at 14.76 nm and the tip is in the second chaotic oscillation state in the hard tapping regime, phase portrait has very complicated structure and the spectrum becomes more broadband as shown in Fig. 2(c). The 3D in Fig. 3(c) is extremely noisy. We conclude that the presence of chaotic tip oscillations deteriorates noticeably the ability of AM-AFM to perform high precision nanometrology.

4. Conclusion

These experimental results indicate that the underlying mechanism of chaos in our experiments during the transition from attractive to repulsive regime is a grazing bifurcation. Specifically, the observed sequence of bifurcations is predicted by the global dynamics of the two-dimensional Nordmark map [12] [13], which is often used to model single degree-of-freedom non-smooth systems. However certain aspects of the experimental results cannot be fully captured by the predictions of the Nordmark map. For instance, the influence of a 1:6 internal resonance between the first and second bending modes is clear in the chaotic spectra. This leads us to conclude that a multiple mode microcantilever model with internal resonance and non-smooth tip-sample interaction forces may be able to capture the observed dynamics.

Finally one may also ask the question about the origin of non-smooth interaction forces in our problem. This is a very relevant question since all theoretical models of atom-atom interactions are based on smooth interaction models. Since we have tried to minimize the effects of liquid menisci in our experiments, we suggest that atomic scale instabilities at the interface of the tip and sample are responsible for non-smooth interaction forces. For instance it is known from molecular dynamics simulations that instabilities occur between atoms at the very end of the tip and the atoms of the sample surface. These instabilities occur at timescales that are extremely small compared to the time scale of tip oscillations leading to effectively non-smooth or discontinuous interaction forces.

Acknowledgments

Financial support of this research by the National Science Foundation (NSF) under Awards 0116414-CMS and No.0409660-CMS is gratefully acknowledged. We would also like to thank Professor R. Reifenberger (Condensed Matter Physics, Purdue University) for helpful discussions.

References


Intermittency in dynamic atomic force microscopy

Ferdinand Jamitzky†‡, Robert W. Stark‡

†Max-Planck-Institut f. Extraterrestrische Physik, Garching, Germany
‡ Center for Nanoscience and Dept. Earth an Environmental Sciences, Ludwig-Maximilians-Universität, Theresienstr. 41, 80333 München, Germany
Email: f.jamitzky@mpe.mpg.de, stark@nanomanipulation.de

Abstract– An atomic force microscope can be viewed as an impact oscillator. These systems are a special class of continuous time dynamical systems which undergo intermittent impact collisions and have dynamical trajectories in state space which are piecewise continuous, however, with discontinuities in the velocities resulting from the collisions of the tip with the surface.

Even if the system is linear in the absence of impacts, such as a freely oscillating AFM cantilever, the overall dynamics of the AFM exhibits a rich variety of behavior because of the non-linearity introduced by the impacts of the tip. In this paper experimental methods have been used in order to investigate the phenomena resulting from the non-linearities induced by the interaction of the tip with the sample surface like hysteresis and subharmonics of the trajectories. In this paper we report about a new chaotic mode of the AFM that has not been observed so far. During the retraction of the tip from the surface the deflection signal shows strong indications of intermittent behavior. The intermittency has been characterized to be either of Type III or on-off intermittency.

Furthermore, understanding the dynamics of the system could help improve the overall system performance by being able to control the AFM in some desired dynamical regime or by preventing the AFM from going into some undesirable regime which would eventually result in system failure.

1. Introduction

Since its invention in 1995 [1] the atomic force microscope (AFM) has become a widely used instrument for surface analysis with nanometer resolution. Various imaging modes including static and dynamic methods have been introduced. The most common techniques are contact mode - a quasi static mode - along with the dynamic measurement modes such as non-contact mode, tapping or intermittent contact mode, and ultrasonic mode[2, 3].

In recent publications [4, 5] it has been shown that the imaging process in the AFM is by no means a linear one, but the existence of non-linearities in the tip-sample interaction can lead to inherent chaotic motions of the tip which finally lead to chaotic distortions of the recorded image. In a experimental work the chaotic modes of the cantilever have been investigated during the approach of a tip to a surface and it has been shown that the motion of the tip undergoes a period-doubling sequence until eventually a chaotic motion has been realized very close to the surface.

It has been often observed that the transition to chaos occurs via the following four known routes [6]:

(i) bifurcation cascade, (ii) intermittency, (iii) crisis and (iv) quasi-periodicity. For the AFM the period-doubling and period-adding cascades have been readily observed experimentally and in numerical simulations.

Furthermore it is well known that the force-distance curves of a tip-surface system show a strong hysteresis and therefore other routes to or from chaos are in principle possible.

In the following we analyze experimental data with tools of non-linear dynamics in order to investigate the complex dynamics of a cantilever in the vicinity of a surface. We provide experimental evidence for an additional route to chaos in the system which has not been observed before and present hints for the origin of the intermittent behavior [7]. The experiment is conducted as a so-called dynamic force spectroscopy experiment where the sample is slowly approached towards the tip. In contrast to experiments where the driving amplitude as a control parameter this corresponds to a typical situation in imaging and nanomanipulation where the amplitude setpoint ratio is used as a control parameter which leads to a variation of the effective tip sample potential.

2. Experimental

Experiments were carried out with an atomic force microscope (Dimension 3100, Veeco, Santa Barbara, CA) was used. Data were recorded with a flexible resolution analog to digital converter (NI PCI-5911, National Instruments, Austin, TX). The microscope was operated in tapping mode.
A cantilever of type CSC 38/Cr-Au Type B has been used which had a resonance frequency of 12.5472 kHz. The driving force had a frequency of 12.5259 kHz and a driving amplitude of 5.1 Volts. The cantilever had a Q-factor of 103. The scan rate was 0.01 Hz and the sampling range in z-direction was 300 nm. A silicon (100) wafer with natural oxide layer was cleaned with ethanol and ultra-pure water.

3. Results and Discussion

The tip was approached and after snap-in to the sample surface it was retracted again. The raw data is shown in figure 2 after reduction by a factor of 10 in order to be visible as single dots. In the first half of the plot positive abscissa values correspond to an approach to the sample surface until the signal breaks down at around sample number 15000. A sharp turning point is visible in the raw data at around 16000 samples. The tip is then retracted from the surface again but still snaps to the surface until sample number 25000 when the tip starts to temporally detach from the surface which can be observed by a sudden widening of the deflection signal. A close-up of the deflection signal is shown in figure 3 where the intermittent behavior can be clearly observed. As the tip is further retracted the quasi-free oscillations prevail and after sample number 30000 the tip is no longer attached to the surface and shows a pure quasi-free motion.

The system clearly exhibits two different kinds of irregular and chaotic motions for the approach and the retraction of the tip. During the approach phase of the tip period-bifurcations have been observed as reported by Jamitzky et al. (2005). It has been shown that the interaction of the tip with the surface potential close to the surface lead to different harmonic distortions and subharmonic motions of the tip which can be characterized by varying embedding dimensions. For example correlation dimension of the attractor of the system ranges from one – far away from the surface to almost three – very close to the surface and show a bifurcation road to chaos. The approach phase will not be further discussed in this paper.

In the following we will focus on the retraction phase of the tip which shows a novel behavior not yet observed in the AFM-system and we will analyze the newly found chaotic regime.

During the retraction phase a hysteresis is observed, as the tip is much longer attached to the surface in contrast to the approach phase but finally snaps out. However the snap-out phase shows a much different behavior compared to the approach phase. Short bursts of a quasi-free motion are alternating with phases where the tip is almost attached to the surface. The tip suddenly snaps away from the surface and starts to oscillate with increasing amplitude until the oscillations become so strong that the tip can be caught again by the surface and stays very close to the surface for a while. The change of the quasi-free and the confined motion of the tip seems to occur rather randomly.

However a clear trend can be observed: the further the tip is retracted from the surface, the longer the quasi-regular phases prevail until the tip is finally detached from the surface and only shows a slight harmonic distortion from a free sine-wave.
The deflection signal in the retraction phase thus shows a strong intermittent behavior. Phases of snap-in of the cantilever to the surface alternate with phases of quasi-regular motion of the tip. But even though the changes between the snap-on and free-motion phases seem to alternate erratically the embedded trajectory clearly shows the existence of an attractor that governs the overall dynamic behavior of the system.

The attractor can be made visible by embedding the time series of the deflection signal into an artificial phase space. For that purpose delay coordinates are used [8, 9], where the signal is delayed for different times and these time series are collected together in a higher dimensional vector which can then be plotted and further analyzed. Here we start with an embedding dimension of three. For short delay times the structure of the attractor is shown in figure 4.

The attractor consists of a spiral structure corresponding to the quasi-free motion of the tip and a point-like attractor which corresponds to the snap-in cantilever. The spiral structure is very pronounced and corresponds to a limit circle of the system which is given by the free motion of the cantilever.

The motion shows a high temporal coherence as can be observed in figure 6. Here the system is embedded using a large delay for the artificial delay coordinates. Even for four times the period of the driving force of the system it shows a very pronounced structure and the attractor decays into several parts.

Fig. 4: Delay coordinate embedding of the deflection signal into an artificial phase space using a short delay time of half a period at time 25500.

The point-like structure corresponds to a fixed-point of the system which is given by the rest position of the tip on the surface. The system can thus oscillate between these two states and does this on the strange attractor that connects these both. When the tip is retracted further, the point-like structure is dissolved until finally the limit circle of the quasi-free motion dominates as shown in figure 5.

Fig. 5: Delay coordinate embedding of the deflection signal into an artificial phase space using a large delay time at time 28800.

Fig. 6: Delay coordinate embedding of the deflection signal into an artificial phase space using a large delay time of 4 times the period of the driving force at time 25500.
Phenomenologically, intermittency is classified in three types depending on the exponent of the power spectrum. In figure 7 the power spectrum is shown for the intermittent phase of the deflection signal. Also shown are spectra with exponent -1 and -2. An exponent of -1 corresponds to type II or type III intermittency while type I intermittency would correspond to an exponent of -1/2. Thus type I intermittency is ruled out.

Fig. 7: Power spectrum of the intermittent phase of the retraction phase. Inserted are spectra corresponding to exponents -1 and -2.

4. Conclusions
The driven cantilever system shows a rich variety of chaotic modes. While period bifurcations have well been observed a route to chaos via intermittency has not been known so far and it is the first time that it is shown experimentally using an AFM. The found intermittency is either of type II or type III, the difference between them is the occurrence in one-dimensional maps like the logistic map [10, 11]. A possible candidate would also be on-off intermittency which shows an exponent of -1. This type of intermittency has been observed by Townsend [12] in the structure of turbulent flows where large eddies alternate with regions of small-scale turbulence. Intermittent behavior has also been observed in the Duffing-system (Brekhovskikh et al [13]) whereas they claim type-I intermittency for the Duffing-system.

An example for intermittency can be nicely observed in the game of pinball. The motion of the ball is almost unpredictable because of the multiple collisions with the walls, however, when the ball gets trapped between a wall and a post the ball oscillates in the “channel” until it rolls down. This type of motion is a good paradigm for the behavior of the tip close to the surface during the retraction phase, as in that case the stiffness of the tip plays an important role in the phases where the tip oscillates quasi-free.

Acknowledgments
Financial support by the German Federal Ministry of Education and Research (BMBF, Grant 03N807) is gratefully acknowledged.

References
[7] Y Pomeau, P Manneville, "Intermittent transition to turbulence in dissipative dynamical systems", Communications in Mathematical Physics, Volume 74, Number 2., 1980
Energy of floating particles aggregates
Vakhtang Putkaradze
Department of Mathematics, Colorado State University, Fort Collins CO 80536
email: putkarad@math.colostate.edu

Abstract—Particles floating on the surface of fluid tend to clump due surface tension interactions (cheerios effect). Similar phenomenon happen at the nanoscale, and mathematical formulation is similar to the millimeter scales. The elevation of the surface is described by the Helmholtz equation with boundary conditions at the particles formed by the contact angle. Due to the weakness of interactions, a high precision numerical calculation is required. We develop a numerical scheme for computing the energy of interactions of such particles. The scheme is based on re-formulating the differential equation as a boundary integral and breaking the integral kernel into the singular (logarithmic and pole) part. The contribution of the singular part is computed analytically, and regular part of the kernel can be approximated with high precision using quadrature formulas. We present the results for simulations of particle interaction for random particle configuration and show that the half-sum of binary interactions is a poor approximation to the energy when the packing of particles becomes dense.

1. Attraction of floating particles: from millimeter to nanometer scales

Particles floating on the surface of fluid tend to clump; this is called cheerios effect after the popular morning cereal. In addition to its entertainment value, the floating particle attraction has been used for the bottom-up self-assembly of objects with the aim of technological and biological applications [1, 2, 3]. For nano-scale applications, one pours out a thin layer of fluid comprising a colloidal solution of nano-particles suspended in water at an interface and allows the water to evaporate. The evaporating fluid film breaks into droplets and drags the particles along, concentrating them at the receding contact line. This effect has been used to deposit nano-particles into prefabricated cavities cut into a substrate, in a process called “directed self-assembly” [3]. Evaporation diminishes the water level until eventually the contact line becomes pinned and the thickness of the film reaches the particle size. At that point, particles which are now only partially submerged in water start to feel each other’s presence due to surface tension.

In view of its potential technological significance, attraction of particles by capillary forces has been an active object of study for many decades. At millimeter scales, particles attract each other through a combination of gravity which deforms the surface so that each particle is downhill from its neighbors and surface tension which resists the deformation. Most previous studies have been devoted to the interaction of two particles [4, 5, 6, 7], implicitly assuming that the interaction of many particles can be obtained as a sum over binary interactions, with, perhaps, corrections by the interactions among triads, tetrads etc. This approach was pioneered by Onsager and Fuoss [8] and has been the cornerstone for the modern theory of mixtures. Two conditions are needed for successful application of this approach. First, the number of foreign inclusions in a fixed volume of fluid must be large enough that statistical averaging is possible, and second, the relative concentration of the inclusions must be small. For example, in the electrolyte solutions originally considered by Onsager and Fuoss, the relative concentration of electrolytes may be assumed to be very small, while the number of electrolyte molecules in any fixed (even relatively small) volume may be assumed large, thereby allowing statistical averaging. Both of these conditions fail for the attraction of nano-particles and cheerios attraction. First of all, concentration of particles is large so maximal possible agglomeration density is achieved, and second, number of particles is relatively small, especially for millimeter-sized particles. In our opinion, the only consistent way to study energies of dense clumps is to make an accurate numerical simulations of particle in various configurations. In this paper, we will summarize the essence of the method we have used for calculations, and present some results. For more detailed information about the results, the reader is referred to [9].

2. Problem set up

The interaction of two particles and numerical scheme has already been described in some details in [10]. We shall refer the reader to that paper for details. However, to make this paper self-contained, we present a short introduction to the methods and resulting integral equations.

It has been shown [4] that surface tension forces are dominating for the micron-size particles, which resulted in the power law for particle interaction. For
much smaller scales (nanometer) which are of interest to us, we can show that in addition to surface tension, van-der-Waals forces must be taken into account [13], for much larger scales (mm), inclusion of gravity is necessary. The interaction potential between two millimeter-size particles floating on water surface and experiencing mutual attraction by surface tension has been studied in great detail, see [7] and references therein. We shall recall this derivation briefly.

Suppose coordinates along the undisturbed surface are chosen as \( x, y \), and the deformation of the surface from equilibrium is given by \( h(x, y) \). As is known, extra pressure caused by surface deformation from the equilibrium \( h \) is \( P_a = \gamma \Delta h \), where \( \gamma \) is the coefficient of surface tension and \( \Delta \) is the two-dimensional Laplace operator. This pressure has to balance the hydrostatic pressure, which is given by \( P_h = \rho gh \). Thus, the deviation of free surface from equilibrium is governed by the Helmhotlz equation

\[
\gamma \Delta h - \rho gh = 0 \quad (1)
\]

We can introduce the capillary length

\[
l_c = \sqrt{\frac{\gamma}{\rho g}}
\]

and re-write equation (1) as

\[
l_c^2 \Delta h - h = 0 \quad (3)
\]

Equation (3) must be supplemented by boundary conditions at the boundary of each particle. We choose to specify contact angle of water-particle-air interface. Exact value of derivative of \( h(x, y) \) with respect to the normal to the water/particle boundary is a nonlinear function of the water level, as the angle of the normal to a sphere relative to the horizontal plane at a given level changes when that level rises or sinks. We have incorporated this nonlinear boundary condition into our scheme and discovered that the result for energy is practically identical to the case where height is computed from the linearization about its equilibrium value. Thus, we shall consider the linearized boundary conditions as it will allow us to compute the energy of large number of interacting particles.

Due to the linearity of equation (3), we shall re-scale the value of the normal derivative of \( h(x, y) \) at the boundary to be 1, as this derivative is assumed to be the same for all particles. This approximation is only valid if we posit that the equilibrium level of water is the same for all particles, and all particles are identical in size, as well as their wetting properties.

Thus, we augment the equation (3) by the following boundary condition on the boundary of each disk \( D_k \)

\[
\frac{\partial h}{\partial n} \big|_{(x, y) \in D_k} = 1. \quad (4)
\]

Earlier work ([13, 10]) demonstrates that for attraction of particles at the nanoscales (\( \sim 50\text{nm} \)), the van-der-Waals force between the fluid and substrate can be linearized about its equilibrium value, which leads exactly to Helmhotlz equations (3) with boundary conditions (4). The value of \( l_c \), computed from available values of physical constants [11], is equal to \( l_c \sim 300 \text{nm} \) for height of the fluid being equal to 50 nm.

These physical considerations show that interaction potentials proportional to the Green’s function for the Helmhotlz operator plays a fundamental role in particle self-assembly across surprisingly many orders of magnitude.

3. The numerical scheme

The technical part of this paper deals with finding an efficient way of solving Helmhotlz differential equation (3) with the boundary conditions (4). The Helmhotlz equation is solved in the outside of the union of \( N \) disks of radius \( R \) with their centers positioned at \( r = c_i = (x_i, y_i) \), \( i = 1, \ldots, N \). An exact solution can be found in the case of one disk; if \( r \) is the polar distance away from the disk, the solution is \( h(r) = l_c K_0(r/l_c) / K_1(r/l_c) \). This exact solution is going to be used for calibration of different numerical schemes.

To put things in perspective, we have used build-in MATLAB routine for solving Helmhotlz equation in a domain. This scheme uses finite elements (triangles) and thus can solve equations in two-dimensional domains of arbitrary shapes. To achieve the accuracy of 0.01 we needed to use about 15000 triangles in the exterior of a disk, the number of triangles grows rapidly with the accuracy required. We note that 0.01 is not sufficient for our purposes; most of the results require computations better by at least two order of magnitude, and finite elements scheme becomes impractical in that case. Instead, we develop a scheme which is based on the boundary integral methods.

Let us introduce Bond number \( B = R/l_c \), and define Green’s function of Helmhotlz operator \( G_B(r, r') = K_0(B|r-r'|/(2\pi B)) \) to be the Green’s function of Helmhotlz operator, with \( K_0(x) \) being the Modified Bessel Function of the second kind. Then, if we denote the strength of ‘point charges’ along the surface of \( k \)-th disk to be \( s_k(\phi_k) \), we have an integral equation for the boundary conditions (4):

\[
\sum_{k=1}^{N} \int_{r' \in D_k} s(\phi_k) \mathbf{n} \cdot \nabla G_B \left( |r - r'| \right) R d\phi_k = 1, \quad (5)
\]

\[
\nabla G_B \left( |r - r'| \right) = \frac{1}{2\pi} K_1 \left( B |r - r'| \right) \frac{r - r'}{|r - r'|} \quad (6)
\]

If we consider equation (5) for \( m \)-th disk, the integral
contributions can be separated in two parts. The first part, which is most singular, comes from the \( m \)-th disk itself, and the regular part comes over summation and integration over all disks with \( k \neq m \). The accuracy of the scheme hinges on the approximation of the singular part of the integral. The idea for this scheme comes from the following expansion for Bessel function \( K_0(z) \) close to \( z = 0 \): \( K_1(z) \sim 1/z + \log(z/2) + \) (regular function of \( z \)). If we assume that \((r,r')\) are measured at the polar angles \((\phi,\phi')\), we can re-write the function \( K_1 \) in (6) as

\[
K_1 \approx \frac{1}{|\phi - \phi'|} + \log(|\sin(\phi - \phi'|/2) + H(\phi, \phi') \quad (7)
\]

where \( H \) is a regular function when \( \phi \to \phi' \). The term resulting from the principal value of integral from the first singular term \( 1/(\phi - \phi') \) results in \( \delta \)-function. Interestingly enough, the integral of the second logarithmic part can be computed exactly:

\[
\int_0^{2\pi} \log(|\sin(\phi - \phi'|)/2) d\phi' = -\pi \log 4. \quad (8)
\]

Once the integrals over the most singular parts are computed exactly, the remaining integral over the regular part is computed using a high order quadrature formula. We have used Simpson’s method; for \( K \) divisions of the disk and \( N \) disks it is accurate up to order \( \sim NK^{-3} \). For reference, we have also computed the integral equation without selecting the singular term and using first-order quadrature. We have computed an accuracy of each method by comparison with exact solution for one disk. Our results are presented on Fig. 1. It is evident that our scheme is superior not only in the decay of error with increase of number of points, but also the pre-factor of the error; the gain in accuracy due to exact description of singular part is about \( 10^3 \).

The resulting potential energy due to disk attraction is proportional to the elevation of the interface from its undisturbed value at the disk boundary. The potential energy for the set of disks with centers positioned at \( c_k, k = 1 \ldots N \) is thus

\[
E[c_1, \ldots, c_N] = \sum_{k=1}^{N} \int_{|r-c_k|=R} h(r(\phi_k)) R d\phi_k. \quad (9)
\]

The force on \( k \)-th disk is computed as minus the gradient of the energy (9) with respect to the position of \( k \)-th disk’s center \( c_k \):

\[
F_k = -\frac{\partial E}{\partial c_k}. \quad (10)
\]

4. Results

What have we achieved by using our numerical scheme? As we see from Figure. 1, only 20 points is sufficient to achieve accuracy of \( 10^{-4} \). Since surface variation caused by mutual attraction of particles is exceedingly small, we need such accuracy to accurately compute the energy of interaction. Suppose we have \( N \) disks with \( K \) points per disk, then the total number of points is \( NK \), and we have to solve a dense linear system of \( NK \) equations with \( NK \) unknowns. A dense system of 2000 equations is plausible to solve on a personal computer, so we can realistically find the energy of 100 disks. The reader has to remember that we are interested in finding statistical results about the energy for many disk distributions, so each solution of linear equations must be accomplished in feasible amount of time. Our results are computed for \( N = 20, 30, 40, 50 \) disks with 20 points per disk and range of Bond numbers \( 1/B = l_i/R = 1.2, \ldots, 10 \). The positions of the disks are chosen at random, from very dilute to highly densely packed configurations.

On Figure 2, we present the results of 5000 numerical experiments, showing the computed energy of interactions \( E \), divided by the sum of binary energies of particles \( E_\Sigma = \frac{1}{2} \sum_{i} kE_0(|r_i - r_k|) \), with \( E_0(r) \) being the energy of particles separated by a distance \( r \). The horizontal axis on this figure is the density \( \rho \), computed as the number of particles, divided by the area of the complex hull containing all the particles. For small values of the density \( \rho \), \( E \approx E_\Sigma \), as expected. However, for high densities we observe large variations of energy \( E \) compared to \( E_\Sigma \), so the theoretical prediction \( E_\Sigma \) is an unacceptable approximation for the energy of interactions. What is interesting, however,
is that energy $E$ is always larger (in absolute value) than $E_\Sigma$. Thus, mutual interaction of particles can lead to the amplification of interaction only. The exact law of particle interaction for an arbitrary configuration of particles will be addressed in future work [9]. Note that the value $\rho = \rho_\ast \sim 0.90$ is the maximal possible packing density for two dimensions, achieved for hexagonal lattice. We enforce the minimum distance between disks to be $0.2R$ to avoid touching of the boundaries of the disks, so maximal density which we can achieve in our simulation is somewhat less than $\rho_\ast$.

Figure 2: The energy of interaction of many densely packed particles $E$ divided by the sum of binary interactions $E_\Sigma$, versus average density $\rho$. Large variations are visible, especially for high densities.

5. Future directions

Our work is the first step towards bridging the gap between continuous and discrete description of particles’ self-aggregation. The method we have described allows to make accurate computations of energy for arbitrary configuration of particles in two dimensions. If there exists an expression for the energy of particles based on density and, perhaps, some other quantity describing particle position, then a continuous equation for density evolution can be found [12]. The big question is whether such an equation for energy exists, and what integral quantities describing particle positions can be used. This question will be addressed in future work [9].

Aknowledgements

I am grateful for fruitful discussions and encouragement of Profs S. R. J. Brueck, D. D. Holm and P. D. Weidman. This work was partially supported by a grant from the National Science Foundation NSF-DMS-0537891.

References

Traveling localized excitations in micro-cantilever arrays

M. Sato† and A. J. Sievers‡

† Graduate School of Natural Science and Technology, Kanazawa University
Kanazawa, Ishikawa 920-1192, Japan
‡Laboratory of Atomic and Solid State Physics, Cornell University
Ithaca, NY 14853-2501, USA
Email: msato@kenroku.kanazawa-u.ac.jp, sievers@ccmr.cornell.edu

Abstract— Micromechanical cantilever arrays provide the opportunity to study the nonlinear excitations of a discrete nonlinear system in real time. Previously we have demonstrated that, even without impurities, localized excitation can be produced and stabilized by means of a chirped driving frequency outside the plane wave vibration band. Such excitations are called intrinsic localized modes (ILMs). More recently, using the same arrays, moving nonlinear excitations have been observed when driving the system at a frequency within the linear dispersion curve range. These traveling ILMs, which are remarkably stable, have been successfully simulated using a nonlinear lumped element lattice model.

1. Introduction

Since it is straightforward to drive a mesoscopic oscillator or a series of such oscillators to large amplitude with modest input power there is renewed interest in the dynamics of discrete nonlinear systems. Linearly coupled vibrating particles have been studied for centuries but the subject of nonlinearly coupled discrete systems has a much more recent history [1]. The dynamics of a 1-D lattice can be modeled with the lumped elements show in Fig. 1(a). A simple mechanical oscillator is coupled to its neighbors with connecting springs. The oscillator spring and/or coupling springs are nonlinear.

In 1988 it was predicted that in addition to plane wave modes a new feature of such a coupled nonlinear system would be the appearance of localized modes [2]. They are called intrinsic localized modes (ILMs) because impurities are not necessary for their creation. The terms discrete breather or lattice soliton have also been used to describe these excitations [3]. Since the appearance of such localized objects only depends on two features, discreteness and nonlinearity, the same localization properties can be expected to occur in many nonlinearly coupled systems.

So far, ILMs have been studied in only a few solid state crystals [4,5], but many man-made lattice structures such as Josephson junction arrays, coupled waveguide arrays, microelectro mechanical systems (MEMS) arrays, and photonic crystals [3]. ILMs have been attracting interest because of their movability and position programability. Future applications could be optical switches and information processors.

We have been examining both by experiment and with simulations the nonlinear excitations of micromechanical cantilever arrays (Fig. 1 (b)) [6,7]. Their fabrication is straightforward and their oscillation frequency is in a convenient range: higher than for coupled pendula, but lower than other man-made lattices such as Josephson junction arrays. A 1D CCD camera can be used to observe the dynamical behavior of such an array. In the past we have succeeded in observing ILM vibration as well as in manipulating its position in the lattice. Here,

![Fig. 1. (a) Coupled ball and spring model. Nonlinear oscillator springs and/or coupling springs are used. (b) Cantilever array. The sample is made from a 300 nm silicon nitride layer. Each cantilever behaves as a nonlinear oscillator coupled together via overhang region. Length and pitch of the cantilevers are about 50 μm. (c) Experimental set up. A line shaped laser beam is focused on the cantilever tips. Reflected beam is imaged on a 1D CCD camera. At large vibrational amplitude the image of those particular cantilevers darkens.](image-url)
we describe a new feature, the observation of moving localized modes obtained under different steady state excitation conditions.

2. Sample and experimental setup

Figure 1(b) illustrates a cantilever array. It is made from a thin silicon nitride film (~300 nm thick) resting on a silicon substrate. Each cantilever has a transverse vibration mode and the coupling between cantilevers is provided by thin nitride overhang region. Resonance frequencies are from 60 to 150 kHz, and the Q factor is ~10000 in vacuum. The number of cantilevers is 152 and their length and pitch is about 50 μm. Both the cantilever itself and the interconnecting overhang have "hard" or "positive" nonlinearity, i.e., the restoring force becomes larger with increasing amplitude so the resonant frequency increases.

The sample is attached to a piezo-electric transducer (PZT) situated in a vacuum chamber to eliminate air damping. The array is shaken up and down uniformly. The vibration of each cantilever is recorded by a 1D-CCD camera as shown in Fig. 1(c). A laser beam is line focused along the cantilever tips, and the reflected beam is imaged on the camera using a lens (not shown). As the amplitude of a cantilever increases its image darkens.

Since the nonlinearity is positive, ILMs can be created above the band of linear resonant frequencies. For the mono-element array shown in Fig. 1(a), the zone boundary mode has the highest linear resonant frequency and is characterized by π phase change between elements. Because the experimental arrangement relies on uniform excitation, such a mode is hard to excite.

To overcome this coupling problem we have prepared a di-element cantilever array, composed of long and short cantilevers, as shown in Fig. 1(b). The acoustic and optic-like dispersion curves (i.e., normal mode frequency as a function of wave number k.) for the di-element array are shown in Fig. 2. The two bands stem from the high and low resonance frequencies of the two cantilevers. The highest frequency mode is now located at k=0 (uniform), and it can be excited by the uniform driver, because of the imbalance in mass for this mode shape. The ILM appears above the upper branch (horizontal thick line in Fig. 2). Since the localized mode is a wave packet, it has a spread in values along the k-axis in the dispersion curve picture, the more localized the excitation the larger the spread in k-space. The localized shape of the ILM and its relation to the linear dispersion curve is confirmed by simulations.

Figure 2 shows that the bandwidth of the upper branch is smaller than the lower branch. This feature is a consequence of the long range interaction between cantilevers produced by the overhang. The narrow bandwidth of the upper branch results in the formation of narrow ILMs. At the same time the large bandwidth of the lower branch implies a larger spacing between adjacent resonant frequencies, since there are only 152/2=76 modes for each branch. Although the two modes that can be most strongly excited are the highest and lowest frequencies at k=0, many of the modes couple to the uniform driving excitation because of the fixed boundary condition. These features mean that we can excite a single mode almost anywhere in the lower branch by simply adjusting the driver frequency. The dotted horizontal line, labeled (b) in Fig. 2, indicates a typical driver frequency location.

3. Experimental results

Figure 3(a) shows an ILM experimental result for the excitation condition (a) identified in Fig. 2. The horizontal white lines are images of cantilevers at rest. The dark region near the center is due to large amplitude cantilevers associated with the ILM. It is obtained by first increasing the driver frequency with respect to the top of the band (up to 1.1% in a few ms), then maintaining a fixed frequency. It is necessary to excite the highest frequency plane wave mode to its modulational instability point so that it can break up into ILMs. Increasing the frequency with time from the highest frequency linear mode is very effective way to obtain a large amplitude. After a number of ILMs are formed, a few remain locked to the driver and grow in strength with time. Such a large amplitude ILM becomes strongly pinned at a lattice site and is very stable. The manipulation of this stable lattice site location has been demonstrated for a stationary ILM [7].
Figure 3(b) shows the experimental results for the excitation condition (b) in Fig. 2. A dark region traveling and reflecting at the boundaries results in a zigzag pattern. Although the lattice has 152 cantilevers, there is a fault induced boundary around site 130. The final state shown in the figure is obtained by carefully tuning the driving frequency. First, it resonantly excites a linear mode. The mode pattern of this resonance is a standing wave because of the boundary condition. Increasing the driver frequency in small steps, causes the amplitude of the standing wave to grow. Similar to the modulation instability observed in case (a) in Fig. 2, the standing wave pattern breaks up into several traveling modes. The initial result appears chaotic. By carefully increasing the driver frequency, often a single traveling mode results. Since the narrow dark region where the excitation level is largest moves through the lattice, it is called a traveling ILM.

4. Numerical Simulations

To simulate the nonlinear array dynamics, the coupled equations of motion for the nonlinear lumped element model is used, namely,

\[ m_i \ddot{x}_i + \frac{m_i}{\tau} \dot{x}_i + k_{20_i} x_i + k_{40_i} x_i^3 + \sum_{j} k_{2i}^{(j)} (2x_i - x_{i+j} - x_{i-j}) + k_{4i}^{(j)} [(x_i - x_{i+1})^3 + (x_i - x_{i-1})^3] = m_i \alpha \cos(\Omega t), \]

where \( m_i \) is the mass, \( k_{20_i} \) and \( k_{40_i} \) are the coefficients of the onsite harmonic and hard quartic potential terms while \( k_{2i}^{(j)} \) and \( k_{4i}^{(j)} \) identify the corresponding intersite terms. More details about the di-element model are given in Ref. [7]. The number of lattice sites is 100 with fixed boundary conditions. Model parameters are listed in the second column of Table II in Ref. [7]. The equations are integrated forward in time using the fourth order Runge-Kutta routine. At each time step of integration a small amplitude noise field is introduced in order to simulate the effects of random vibrational noise. Figure 4(a) shows simulation results obtained with this model. Comparing the experiment shown in Fig. 3(b) with the simulation in Fig. 4(a) shows a remarkably similar time dependence. The only difference is that the two ILM speeds are not quite the same. This difference occurs because the slope
of the model dispersion curve at the driver frequency is somewhat larger than that observed in experiment. The difference comes about because the driver frequency is 10.6 kHz lower than the top of the lower branch in the simulation, while it is 5.8 kHz from the top for the experiment. A carefully tuning of the driver frequency in the simulation, with sufficient resolution compared to the frequency separation between neighboring plane wave modes, is required to generate a traveling ILM. This initial condition agrees with that previously used in the experiments.

Figure 4(b) presents a simulation result with the identical starting condition as that shown in Fig. 4(a), but now the nonlinear terms have been removed so the array is harmonic. The initial localized wave at \( t=0 \) spreads into many waves with different velocities with increasing time. It is clear that the localized mode envelope is coming apart after only one cycle through the lattice and can no longer be identified after three cycles.

5. Discussion and Conclusions

The spacial localization of the traveling ILM shown in Fig. 3(b) is similar to that found for the stationary ILM, in Fig. 3(a). The simulations show that both types of modes can be described by an envelope modulating a carrier wave, but in the latter case the \( \omega-k \) display appears as a line tangent to the dispersion curve at the driver frequency position. Such a situation is represented by the feature near the dotted line (b) in Fig. 2. The speed of the traveling mode can be estimated from experiment, and good agreement is found with the corresponding tangent line. It is to be expected that all such running ILMs can be represented in an \( \omega-k \) picture as tangents to the linear dispersion curve, because the dispersion curve identifies the positions of available plane wave modes. The existence of such a traveling ILM depends on a balance between the nonlinear and dispersive effects. Here, the negative curvature of the dispersion balances the positive nonlinearity as represented by the two arrows in Fig. 2. This balance is reminiscent of the soliton stability condition in a continuous nonlinear medium; but, unlike that case, where two solitons can pass though each other with only a phase change, for the lattice problem, in the absence of a driver, energy is transferred between two ILMs during their collision.

These experiments demonstrate that micromechanical cantilever arrays provide a readily accessible technique for exploring the properties of localized excitations in a nonlinear lattice. Two kinds of nonlinear localized excitations have been generated: stationary ILMs in the region outside of the plane wave spectrum and traveling ILMs, which appear in the plane wave spectrum. Choosing the driving frequency to occur at a particular dispersion curve frequency determines which kind of localized steady state excitation profile will appear. The good agreement between experiment and simulations demonstrates that the nonlinear lumped element model captures the important physical signatures that have been observed experimentally.

Acknowledgments

MS would like to thank Kenichi Maruno for discussions about solitons. This work was supported in past by NSF DMR 0301035 and by DOE DE-FG02-04ER46154.

References

Suppression of subharmonics and chaos in tapping mode atomic force microscopy using time delayed feedback control

Kohei YAMASUE † and Takashi HIKIHARA †

†Department of Electrical Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615-8510, Japan
Email: yamasue@dove.kuee.kyoto-u.ac.jp

Abstract—Nonlinear oscillation of cantilever sensors in the tapping-mode atomic force microscopy (AFM) has been extensively studied because of its close relation to imaging quality of the tapping-mode AFM. In particular, a recent experimental study by Hu et al. has suggested that the imaging quality is significantly deteriorated due to subharmonics and chaos generated by grazing bifurcation. In this paper, we propose elimination of the subharmonics and chaotic oscillation in a grazing regime using the time delayed feedback control. Possibility of its application is numerically discussed based on stability analysis of cantilever oscillation under the time delayed feedback control.

1. Introduction

The atomic force microscopy (AFM) [1] is nowadays widely utilized for probing nanoscale structures [2]. Among its operation modes, the dynamic mode AFM allows us to investigate material surfaces in nanometer resolution without damaging samples [3, 4]. A wide variety of samples have been observed including semiconducting [5, 6, 7], organic [8], and biological samples even in liquids [9, 10].

A key device of the dynamic mode AFM is a vibrating cantilever sensor to detect a tiny interaction force between a sharp tip manufactured at the free end of the cantilever and a sample surface facing the tip. A topography of the sample surface is acquired by raster scan of the surface with keeping the vibration or resonance frequency of the cantilever constant. However, a variety of nonlinear phenomena have been predicted theoretically and observed experimentally [11, 12, 13, 14, 15]. In particular, nonlinear dynamics of cantilevers in the tapping-mode AFM [3, 16] has been focused on so far due to a highly nonlinear tip-sample interaction force in its operating range [17, 18]. A bistable behavior of the tip oscillation occurs near sample surfaces [19, 13, 14, 17, 18]. Resulting jumping and hysteresis phenomena cause sudden and discontinuous transition of imaging characteristics [17]. In addition, it was reported that the cantilever exhibits subharmonic oscillation, period-doubling bifurcation [20, 21], and chaotic oscillations [22, 23, 24, 25]. Grazing bifurcation in transition from non-contact to contact regime was also predicted numerically by van de Water et al. [26] and recently suggested experimentally by Hu et al. [27]. Since the operating range of the tapping-mode AFM is possibly limited by non-periodic and irregular motion of the cantilever sensors, control techniques for cantilever oscillations should be developed to improve performance of the tapping-mode AFM. Some research groups have, therefore, proposed application of control methods [22, 23, 28] and the authors have also proposed the application of time delayed feedback control [29]. The authors have numerically shown that the application of the control method allows us to extend an operating range of the tapping-mode AFM and accelerate its scanning rate without reducing force sensitivity. Possibility of its application was, however, only discussed in a bistable regime in Ref. [29].

In this paper, we numerically discuss the stabilization of cantilever sensor in a grazing regime using the time delayed feedback control. The time delayed feedback control was originally proposed by Pyragas to stabilize unstable periodic orbits embedded in chaotic attractors [30]. This control method has been extensively studied in the field of nonlinear dynamics [31]. A significant feature is that the control method only employs an error signal between the current and past output of a nonlinear system. The control method, therefore, can be applied without any identifica-
Figure 2: Bifurcation diagram of cantilever oscillation in grazing regime. Horizontal axis corresponds to tip-sample distance monotonously decreased from $\alpha = 27$ to $\alpha = 7$. Vertical axis denotes displacement sampled at the same phase for each period of driving force. We can see periodic oscillations including subharmonic components and chaotic oscillation.

This paper is organized as follows. Section 2 describes a mathematical model of vibrating cantilevers near sample surfaces. The model was originally proposed by Ashhab et al. [23]. An additional control input was applied to this model based on the time delayed feedback control [29]. Section 3 presents a possibility of extending the operating range of tapping-mode AFM. The stability of target periodic oscillation is numerically estimated based on the difference differential equation introduced in Sec. 2.

2. Model of cantilever near sample surfaces

A schematic diagram of the tapping-mode AFM is shown in Fig. 1. The cantilever is externally forced at or near its mechanical resonant frequency during measurement. When the cantilever tip approaches to a sample surface, the resonance frequency is changed depending on a tip-sample interaction force governed by their distance. A topography of the surface is acquired by raster scan of the surface with keeping the resonance frequency constant. Shift of the resonance frequency is detected by measuring amplitude of the vibration in the tapping-mode AFM [3, 16]. The vibration is typically measured by the optical lever method [32].

When the tip-sample interaction force is approximated by a potential function called the Lennard-Jones potential, the first mode vibration of the cantilever is modeled by the following differential equation [23]:

$$\begin{align*}
\dot{x} &= y \\
\dot{y} &= -x - \frac{d}{(\alpha + x)^2} + \frac{\Sigma d}{30(\alpha + x)^8} + \varepsilon(\Gamma \cos \Omega t - \Delta y) + u
\end{align*}$$

where $\alpha$ denotes the equilibrium point of the tip, when only the gravity acts on it. $(x, y)$ is the displacement and velocity of the tip. $(\Omega, \Gamma)$ denotes the frequency and amplitude of the sinusoidal external force driving cantilever. $\Sigma$ is diameter of a molecular composing sample and tip. $d = 4/27$ is a constant. $\varepsilon$ is a small parameter. Equation (1) is dimensionless. It is shown that the cantilever oscillation can be chaotic by applying the Melnikov method to Eq. (1) under $u(t) = 0$ [23].

As firstly proposed by Pyragas, continuous control input $u(t)$ stabilizing unstable periodic orbits is given by the error between the current output and the past one as follows [30]:

$$u = K[g(x_t, y_t) - g(x, y)],$$

(2)

where $\tau$ and $K$ denote control parameters called time delay and feedback gain, respectively. $g(x, y) = g(x(t), y(t))$ and $g(x_t, y_t) = g(x(t-\tau), y(t-\tau))$ imply a scalar output signal measured at the current time $t$ and the past time $t-\tau$, respectively. The control input (2) only depends on the output signal. The control method, therefore, can be implemented to experimental systems without any models of cantilevers nor detailed analyses of underlying dynamics. The control method has been successfully applied to various experimental systems including the magneto-elastic beam [33]. The magneto-elastic beam system [34] and the cantilever located near surfaces have a similar dynamical structure characterized by an elastic beam sinusoidally forced under two-well potential, although there is a difference that dimension of the latter system is so much smaller. The time delay $\tau$ and feedback gain $K$ are important control parameters that governs its control performance. The time delay $\tau$ is adjusted to the period of a target unstable periodic orbit we intend to stabilize in a chaotic attractor. The control input, therefore, converges to null after the controlled system is stabilized to the target orbit. Stability of the target orbit largely depends on the feedback gain.

We hereafter investigate the stability of cantilever oscillations under the time delayed feedback control. The cantilever has the damping coefficient $\Delta = 0.1$. The frequency of driving force is fixed at $\Omega = 0.98$ just below the mechanical resonance frequency, corresponding to one of the experimental setups by Hu et al. [27]. The remaining parameters are set as $\Sigma = 0.3$ and $\varepsilon = 0.1$ based on the numerical result in Ref. [24]. $\alpha$ is set in the regime where the tip of cantilever grazes a sample surface. If the velocity of oscillation is estimated as an output of the nonlinear system (1), the control signal $u(t)$ is given as follows:

$$u(t) = K[y(t-\tau) - y(t)].$$

(3)

The time delay $\tau$ is adjusted to $2\pi/\Omega$ to stabilize an orbit with the same frequency as the external force. Stabilization of this orbit is necessary for the tapping-mode AFM. The operating range may be limited by non-periodic and irregular oscillations caused by chaos. In addition, the force sensitivity of tapping-mode AFM is possibly decreased in
the standard device configuration using lock-in amplifiers or RMS-DC converters with band-pass filters. This is because subharmonics and chaos make frequency spectrum of the oscillation widely spread, while only a particular frequency component is detected in the standard configuration.

3. Extension of operating range

An experimental study by Hu et al. has demonstrated chaotic oscillations of cantilevers in a tapping-mode AFM [27]. They suggested the chaotic oscillation is caused by grazing bifurcation, which occurs when the cantilever tip just begins to hit a surface.

We here numerically show that the time delayed feedback control has an ability to stabilize subharmonic and chaotic oscillations in a grazing regime. When the cantilever approaches the surface, a chaotic oscillation is generated. A corresponding bifurcation diagram is shown in Fig. 2, which is obtained by monotonically decreasing the tip-sample distance from $\alpha = 27$ to $\alpha = 7$. The vertical axis shows displacement sampled at the same phase for each period of driving force. One can see that period-$2\pi/\Omega$ oscillation is bifurcated to chaotic oscillation and periodic oscillations including subharmonic components. Figure 3 shows a chaotic attractor for $\alpha = 20$ and an unstable periodic orbit embedded in it. The chaotic attractor and embedded unstable periodic orbit are shown by gray and black line, respectively. We should stabilize the orbit for operation of the tapping-mode AFM. The orbit is hereafter target orbit. The target orbit is stabilized when we adjust the time delay $\tau$ to $2\pi/\Omega$ as shown in Fig. 4. The chaotic oscillation of tip is changed to the target periodic one after the activation of control. The feedback gain is adjusted to $K = 0.075$. The time of activation is pointed by an arrow in Fig. 4(b). We confirm that the displacement of tip in Fig. 4(a) is changed from chaotic to a periodic oscillation. Corresponding to this change, control input in Fig. 4(b) converges to null signal after the activation. This null signal implies that the chaotic oscillation is eliminated by stabilizing the target unstable periodic orbit embedded in the chaotic attractor.

Note that the control input changes only the stability of the target orbit. The control input does not modify any system parameters in contrast to the feedback control proposed in Refs. [22, 23]. It is, therefore, possible to extend the operating range of the tapping-mode AFM. Figure 5 shows numerically estimated spectral radius of the target orbit for various feedback gain. The spectral radius is the modulus of the characteristic multiplier that has maximal modulus among all characteristic multipliers of the target orbit. The target orbit is stable, if the spectral radius is less than unity. It is observed that smaller spectral radius is allowed under control input. In particular, the spectral radius for $K = 0.1$ is less than unity in the whole range we investigate. This implies that the tapping-mode AFM can be stably operated in the grazing regime using the time delayed feedback control.

4. Conclusion

In this paper, we numerically discussed the stabilization of cantilever sensors in the tapping-mode AFM using the time delayed feedback control. Possibility of the stabilization in the grazing regime was shown by stability analysis of the target periodic oscillation. We can, therefore, extend the operating range of the tapping-mode AFM to the grazing regime where the subharmonics and chaotic oscillation may be produced. This implies that intrinsic dynamics behind unstable oscillations including chaos can be utilized.
for measurement by applying the time delayed feedback control.

Acknowledgments

The authors would like to thank Professor Igor Mezić, University of California, Santa Barbara, for his kind acceptance of our use of model of dynamic mode AFM. Our research is motivated by theoretical and numerical results of his research group. The authors also thank Professor Kazumi Matsushige, Professor Hirofumi Yamada, and all other members of electronic material science and engineering laboratory, Kyoto University, for fruitful discussions and valuable suggestions on actual devices and their operation. The authors would like to appreciate Dr. Robert Walter Stark, Dr. Ferdinand Jamitzky, and Dr. Javier Rubio-Sierra, Ludwig-Maximilians-Universität München for their precious comments on nonlinear dynamics and control of cantilevers in the AFM. This research is partly supported by the Ministry of Education, Culture, Sports, Science and Technology in Japan, The 21st Century COE Program No. 114213201 and also supported by Iketani Science and Technology Foundation, Grant No. 0181067A.

References

Lagrangian stability of viscous flows

Yodai Watanabe†‡

†National Institute of Informatics, Research Organization of Information and Systems, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan
‡Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, 2-1 Hirosawa, Wako-shi, Saitama 351-0198, Japan

Abstract—A differential geometrical formulation of the motion of an incompressible viscous fluid is presented. The geodesic equation on a manifold is extended with an additional term consisting of an endomorphism of the tangent space, and a class of curves defined by the resultant equation is introduced. Based on this extension, the motion of an incompressible viscous fluid is formulated as curves of this class in a diffeomorphism group, and the expression for the variational equation of the curves is derived. The expression is then shown to coincide with the governing equations for the Lagrangian displacement, which interprets the physical meaning of variation vector fields of the curves. The variational equation is reduced to a more simplified form, and by solving the simplified equation, Lagrangian stability of a plane parallel flow is investigated. 1

1. Introduction

The motion of an incompressible fluid of a compact domain (manifold) can be described as curves in the volume-preserving diffeomorphism group \( \mathcal{D}_\mu(M) \) of the domain \( M \). In particular, if the fluid is inviscid, then its motion is formulated as geodesics in \( \mathcal{D}_\mu(M) \)[1]. This is an example of the fact that the motion of a dynamical system with a Lie–Poisson structure can be formulated as geodesics in a Lie group with an appropriate metric[9]. (For other examples of the Lie–Poisson systems, see [2, 14, 15] etc.) In this formulation, the original dynamical system is defined on a Lie algebra, while the geodesic equation associated with the system is defined on a Lie group. In the case of an incompressible fluid, for instance, the Lie algebra corresponds to the space of velocity fields, while the Lie group corresponds to the space of particle configurations. Therefore stability of geodesics in \( \mathcal{D}_\mu(M) \) corresponds to Lagrangian stability of fluid flows, i.e. stability of the motion of fluid particles.

The Lie–Poisson structure is a very special one; therefore it is rare that a system of interest has a Lie–Poisson structure and is applicable of the above formulation. Thus, in this paper, we extend the formulation so that it will be applicable to more general systems including incompressible viscous fluids. For this purpose, in section 2, we extend the geodesic equation on a manifold with an additional term consisting of an endomorphism of the tangent space, and introduce a class of curves, called \( \tau \)-geodesics, which are defined by the extended form of the geodesic equation. This extension allows us to formulate the motion of an incompressible viscous fluid as \( \tau \)-geodesics in the diffeomorphism group \( \mathcal{D}_\mu(M) \) (see [17]). In section 3, we derive the expression for the variational equation of \( \tau \)-geodesics in \( \mathcal{D}_\mu(M) \), and then show that the expression is equivalent to the governing equations for the Lagrangian displacement.

In section 4, we study Lagrangian chaos in a plane parallel flow. Section 5 is devoted to conclusion of this paper.

\section{Extension of the geodesic equation}

This and the next sections provide a review of definitions and results concerning \( \tau \)-geodesics and a geometrical formulation of viscous flows. See [17] for details.

Let \( M \) be a smooth manifold (i.e. a differentiable manifold of class \( C^\infty \)) with an affine connection \( \nabla \). A smooth curve in \( M \) is called a geodesic if it satisfies the following differential equation, called the geodesic equation,

\[ \nabla_x \dot{x} = 0, \tag{1} \]

where \( \dot{x} = dx/dt \). In this section, we extend the geodesic equation with an additional term consisting of an endomorphism of the tangent space.

\textbf{Definition 1} A \( \tau \)-affine manifold is a triplet \((M, \nabla, \tau)\) such that \( M \) is a smooth manifold, \( \nabla \) is an affine connection on \( M \), and \( \tau \) is a smooth tensor field of type \((1, 1)\) on \( M \).

Note that, for each \( x \in M \), \( \tau_x \) can be regarded as an endomorphism of the tangent space \( T_x M \). The endomorphism will be denoted by \( \tau(x) \) (or \( \tau_x \) at \( x \in M \)). We can now introduce the operator \( \tilde{\nabla} : TM \times \Gamma(TM) \to TM \) by writing

\[ \tilde{\nabla}_X Y = \nabla_X Y + \tau(X), \tag{2} \]

where \( X \) and \( Y \) are vector fields on \( M \), and \( \Gamma(TM) \) is the set of all the sections of \( TM \). This operator is no longer a connection, but it defines a curve in \( M \) as an affine connection defines a geodesic. Such a curve will be called a \( \tau \)-geodesic; more precisely,
Definition 2 Let \((M, \nabla, \tau)\) be a \(\tau\)-affine manifold. A smooth curve \(x_t, a < t < b,\) in \(M\) is called a \(\tau\)-geodesic if it satisfies the following ordinary differential equation, called the \(\tau\)-geodesic equation,
\[
\ddot{x} + \tau(\dot{x}) = 0 \quad \text{for } a < t < b.
\]
(3)
Let \(T\) and \(R\) be the torsion and the curvature tensor fields on a smooth manifold \(M\) with an affine connection \(\nabla\). A vector field \(J_t\) along a geodesic \(x_t\) in \(M\) is called a \(\tau\)-Jacobi field if it satisfies the following ordinary linear differential equation, called the \(\tau\)-Jacobi equation,
\[
\ddot{\xi} + \tau(\dot{\xi}) = 0 \quad \text{for } a < t < b.
\]
(4)
The following is the analogue of this definition.

Definition 3 Let \((M, \nabla, \tau)\) be a \(\tau\)-affine manifold. A vector field \(J_t\) along a \(\tau\)-geodesic \(x_t\) in \(M\) is called a \(\tau\)-Jacobi field if it satisfies the following ordinary linear differential equation, called the \(\tau\)-Jacobi equation,
\[
\nabla_t \nabla_t J + \nabla_t T(J, \dot{x}) + R(J, \dot{x}) J_t = 0.
\]
(5)
where \(T_t\) is defined by
\[
T_t(X, Y) = \nabla_X \tau(Y) - \nabla_Y \tau(X) - \tau([X, Y]).
\]
(6)
It readily follows from the definition of \(T_t\) that
\[
T_t(X, Y) = -T_t(Y, X), \quad T_t(fX, Y) = fT_t(X, Y),
\]
(7)
and so \(T_t\) is an antisymmetric tensor field. It is therefore clear that the \(\tau\)-Jacobi equation (5) is a second-order differential equation for \(J_t\). See [17] for a geometrical interpretation of a \(\tau\)-Jacobi field.

Let \(G\) be a Lie group and \(g\) its Lie algebra. For \(X \in g\), let \(X_e\) denote the right-invariant vector field on \(G\) such that \((X_e)_x = X\), where \(e\) is the identity element of \(G\). Now we examine a Lie group \(G\) with a right-invariant affine connection \(\nabla\) and tensor field \(\tau\). Let \(Y_t\) be a vector field along a curve \(x_t\) in \(G\), and let \(U_t\) and \(V_t\) be the curves in \(G = T_t G\) defined by \(U_t = d_x R_{x_t}^{-1}(X_t)\) and \(V_t = d_x R_{x_t}^{-1}(Y_t)\), respectively. Then,
\[
\left( \nabla_t Y_{x_t} \right)_t = d_x R_{x_t} \left( \frac{d}{dt} V + \left( \nabla_t V_{x_t} \right)_t \right).
\]
(8)
This equation is useful for reducing a differential equation on \(G\) described in terms of \(\nabla\) to that on \(g\).

Proposition 4 ([16, 17]) Let \((G, \nabla, \tau)\) be a \(\tau\)-affine manifold such that \(G\) is a Lie group, and \(\nabla\) and \(\tau\) are both right-invariant. Let \(x_t\) be a curve in \(G\), and let \(U_t\) be the curve in \(g = T_t G\) defined by \(U_t = d_x R_{x_t}^{-1}(X_t)\). Then \(x_t\) is a \(\tau\)-geodesic if and only if \(U_t\) satisfies
\[
\frac{d}{dt} U_t + \left( \nabla_t U_{x_t} \right)_t + \tau_t(U) = 0.
\]
(9)
Proof. See [17]. \(\square\)

Similarly, the \(\tau\)-Jacobi equation on \(G\) reduces to a differential equation on \(g\) which has a more convenient and suggestive form.

Theorem 5 ([17]) Let \((G, \nabla, \tau)\) be a \(\tau\)-affine manifold such that \(G\) is a Lie group, and \(\nabla\) and \(\tau\) are both right-invariant. Let \(J_t\) be a vector field along a \(\tau\)-geodesic \(x_t\), and let \(U_t\) and \(V_t\) be the curves in \(g = T_t G\) defined by \(U_t = d_x R_{x_t}^{-1}(X_t)\) and \(V_t = d_x R_{x_t}^{-1}(Y_t)\), respectively. Then \(J_t\) is a \(\tau\)-Jacobi field along \(x_t\) if and only if \(V_t\) satisfies
\[
\frac{d}{dt} U_t' + \left( \nabla_t U_{x_t} \right)_t + \left( \nabla_t U_{x_t} \right)_t + \tau_t(U) = 0,
\]
(10)
where we have defined
\[
U_t' = \frac{d}{dt} U + \left( \left( U_t, V_t \right)_t \right).
\]
(11)
Proof. See [17]. \(\square\)

3. Lagrangian displacement and Lagrangian stability

Let \(D_\rho(M)\) be the group of volume-preserving smooth diffeomorphisms of \(M\). The group multiplication in \(D_\rho(M)\) is the composition of mappings. Group \(D_\rho(M)\) with the inverse limit topology is an infinite-dimensional Lie group called an inverse limit Hilbert (ILH) Lie group. We thus assume a smooth manifold structure and smooth group operations on \(D_\rho(M)\). (For details, see [6, 12]). The Lie algebra of \(D_\rho(M)\) is the set \(\mathfrak{X}_\rho(M)\) of smooth divergence-free vector fields on \(M\). Elements of \(\mathfrak{X}_\rho(M)\) will be denoted as \(U = u, V = v\) etc., where \(u, v\) etc. are vector fields on \(M\).

To consider \(\tau\)-geodesics in \(D_\rho(M)\), we now define a right-invariant affine connection and tensor field of type \((1, 1)\) on \(D_\rho(M)\). Let \(\langle \cdot, \cdot \rangle\) be the inner product of \(\mathfrak{X}_\rho(M)\) defined by
\[
\langle U, V \rangle = \int_M u \cdot v \, dx
\]
(12)
and \(\tau_e(\cdot)\) be the endomorphism of \(\mathfrak{X}_\rho(M)\) defined by
\[
\tau_e(U) = -\frac{1}{R} \Delta u.
\]
(13)
The inner product \(\langle \cdot, \cdot \rangle\) and the endomorphism \(\tau_e(\cdot)\) are extended by the right translation to induce the right-invariant metric \(\langle \cdot, \cdot \rangle\) and the right-invariant tensor field \(\tau(\cdot)\) of type \((1, 1)\), respectively. Let \(\nabla\) be the Riemannian connection compatible with the metric \(\langle \cdot, \cdot \rangle\). Then, we can show the following claim (see [17] for details).

Claim 6 ([17]) Let \((D_\rho(M), \nabla, \tau)\) be as above. Then the \(\tau\)-geodesic equation on \(D_\rho(M)\) is equivalent to the Navier-Stokes equations.

Let \(\xi_t\) be a \(\tau\)-geodesic in \(D_\rho(M)\) and \(J_t\) a \(\tau\)-Jacobi field along \(\xi_t\). Set \(U_t = d_x R_{\xi_t}^{-1}(\xi_t)\) and \(v_t = d_x R_{\xi_t}^{-1}(J_t)\). It then follows that \(U_t\) is a solution of the Navier–Stokes equations,
and \( \mathbf{v} \) is governed by the \( \tau \)-Jacobi equation (10) with (11). Hence, on introducing a velocity field \( \mathbf{u}' \) to represent \( \mathbf{U}' \) in (11), it follows that

\[
\frac{\partial}{\partial t} \mathbf{u}' + (\mathbf{u}' \cdot \nabla) \mathbf{U} + (\mathbf{U} \cdot \nabla) \mathbf{u}' = -\nabla p' + \frac{1}{\mu} \Delta \mathbf{u}',
\]

\[
\text{div} \mathbf{u}' = 0 \tag{14}
\]

and

\[
\mathbf{u}' = \frac{\partial}{\partial t} \mathbf{v} + (\mathbf{U} \cdot \nabla) \mathbf{v} - (\mathbf{v} \cdot \nabla) \mathbf{U}, \tag{15}
\]

respectively. These equations completely coincide with the governing equations for the Lagrangian displacement (for details of the Lagrangian displacement, see [3, 4]). Consequently, we obtain the following claim.

**Claim 7 ([17])** Let \( \mathcal{D}_s(M), \nabla, \tau \) be as above. Then the \( \tau \)-Jacobi equation on \( \mathcal{D}_s(M) \) is equivalent to the governing equations for the Lagrangian displacement.

In general, Lagrangian stability (i.e. stability of the motion of fluid particles) does not coincide with Eulerian stability (i.e. stability of the velocity field). For instance, there is a flow which is stable in the latter sense but unstable in the former sense (see e.g., [10]). In this paper, we call such a flow Lagrangian chaotic. It should be mentioned that Lagrangian chaotic flows are not only of theoretical interest but also of practical importance in mixing viscous fluids (see [13]). As long as we consider the linear stability, equations (14) and (15) are sufficient for investigating the Eulerian and Lagrangian stabilities, because evolution of the Lagrangian displacement \( \mathbf{v} \) and the perturbation quantities \( \mathbf{u}' \) and \( p' \) are completely described by these equations. Since equations (14) are linear in \( \mathbf{u}' \), they have the trivial solution \( \mathbf{u}' = 0 \), for which the governing equations for the Lagrangian displacement reduce to the simplified form

\[
\frac{\partial}{\partial t} \mathbf{v} + (\mathbf{U} \cdot \nabla) \mathbf{v} - (\mathbf{v} \cdot \nabla) \mathbf{U} = 0. \tag{16}
\]

Note that the condition \( \mathbf{u}' = 0 \) implies that there is no perturbation to the basic flow \( \mathbf{U} \). Hence, the above equation describes evolution of the (infinitesimal) distances between fluid particles advected by the unperturbed flow \( \mathbf{U} \). It thus follows that we can use this equation to examine whether individual fluid particles initially placed in a sufficiently small region diverge from each other or not.

Although equation (16) is much simpler than the original equations, it is useful for studying Lagrangian chaotic flows in particular when the basic flow \( \mathbf{U} \) is steady, i.e. when \( \mathbf{U} \) is independent of \( t \). In this case, the system of the original linear partial differential equations (14) and (15) has coefficients independent of \( t \). Therefore we may separate the variables and take normal modes of the form

\[
\mathbf{v}(t, \mathbf{x}) = e^{st} \hat{\mathbf{v}}_s(\mathbf{x}), \quad \mathbf{u}'(t, \mathbf{x}) = e^{st} \hat{\mathbf{u}}'_s(\mathbf{x}), \quad \rho'(t, \mathbf{x}) = e^{st} \hat{\rho}'_s(\mathbf{x}), \tag{17}
\]

Here eigenvalue \( s \) and corresponding eigenfunctions \( \hat{\mathbf{v}}_s, \hat{\mathbf{u}}'_s \) and \( \hat{\rho}'_s \) can be found in principle by solving the resultant equations

\[
s\hat{\mathbf{v}}_s + (\hat{\mathbf{u}}'_s \cdot \nabla) \mathbf{U} + (\mathbf{U} \cdot \nabla) \hat{\mathbf{v}}_s = -\nabla \hat{\rho}'_s + \frac{1}{\mu} \Delta \hat{\mathbf{u}}'_s, \quad \text{div} \hat{\mathbf{u}}'_s = 0, \tag{18}
\]

\[
s\hat{\mathbf{v}}'_s + (\hat{\mathbf{u}}'_s \cdot \nabla) \hat{\mathbf{v}}_s - (\hat{\mathbf{v}}'_s \cdot \nabla) \mathbf{U} = \hat{\mathbf{u}}'_s \tag{19}
\]

in \( M \), with the resultant boundary conditions on \( \partial M \). We mention that the separation of the variable \( t \) can be justified in many cases by use of a Laplace transform. Suppose further that the basic flow \( \mathbf{U} \) is Lagrangian chaotic, i.e. stable in the Eulerian sense but unstable in the Lagrangian sense. It then follows from the Lagrangian instability of \( \mathbf{U} \) that there exists at least one eigenvalue \( s \) such that \( \text{Re}(s) > 0 \). However it also follows from the Eulerian stability of \( \mathbf{U} \) that the eigenfunctions \( \hat{\mathbf{u}}'_s \) and \( \hat{\rho}'_s \) vanish for all eigenvalues \( s \) such that \( \text{Re}(s) > 0 \). Hence

\[
s\hat{\mathbf{v}}'_s + (\hat{\mathbf{u}}'_s \cdot \nabla) \hat{\mathbf{v}}_s - (\hat{\mathbf{v}}'_s \cdot \nabla) \mathbf{U} = 0 \quad \text{if Re}(s) > 0, \tag{21}
\]

and so we obtain the following claim.

**Claim 8 ([17])** Suppose that a basic flow \( \mathbf{U} \) is steady and Lagrangian chaotic. Assume that the general solution of equations (14) and (15) can be expressed as a linear superposition of normal modes of the form

\[
\mathbf{v}(t, \mathbf{x}) = e^{st} \hat{\mathbf{v}}_s(\mathbf{x}). \tag{22}
\]

Then all unstable modes satisfy equation (16).

This claim shows that the condition for Lagrangian instability of a steady laminar flow can be found by solving equation (16).

### 4. Lagrangian stability of a plane parallel flow

In this section, we study Lagrangian chaos in a plane parallel flow. A plane parallel flow is that with velocity components and pressure of the form

\[
\mathbf{U} = (U(z), 0, 0), \quad \rho = \rho(z). \tag{23}
\]

The flow is assumed to be bounded by the two rigid planes \( z = 1 \) and \( z = -1 \) (Figure 1). For an inviscid fluid, it can be verified, by substituting the equations above into the Euler equations, that \( \mathbf{U} \) can be an arbitrary function of \( z \). However, for a viscous fluid, it can also be shown, by substituting them into the Navier-Stokes equations, that \( \mathbf{U} \) must be of the form

\[
\mathbf{U} = az^2 + bz + c, \tag{24}
\]

which describes the Couette-Poiseuille flow. The constants \( a, b \) and \( c \) are determined by the pressure gradient and boundary conditions.

Let us now consider Lagrangian chaos in a plane parallel flow. As was shown in the previous section, Lagrangian
chaotic flows can be found by examining unstable modes of equation (16). Taking normal modes of the form
\[ v = (u(z)e^{-ikx+iby}, \nu(z)e^{-ikx+iby}, w(z)e^{-ikx+iby}), \]
we see at once from equation (16) that
\[ su + ikUu - wU' = 0, \]
\[ sv + ikUv = 0, \]
\[ sw + ikUw = 0. \]
It readily follows that \( \text{Re}(s) > 0 \) only when \( u(z) = v(z) = w(z) = 0 \). Consequently, we have the following claim.

**Claim 9** A plane parallel flow cannot be Lagrangian chaotic; that is, a plane parallel flow is Lagrangian unstable only when it is Eulerian unstable.

### 5. Conclusion

This paper is concerned with Lagrangian stability of viscous flows. First we extended the geodesic equation on a manifold with an additional term consisting of an endomorphism of the tangent space, and introduce a class of curves defined by the resultant equation. Based on this extension, we formulated the motion of an incompressible viscous fluid as curves of this class in a di
tension, we formulated the motion of an incompressible
viscous fluid as curves of this class in a di
tension, we formulated the motion of an incompressible
viscous fluid as curves of this class in a di
tension, we formulated the motion of an incompressible
class of curves defined by the resultant equation. Based on this ex
tension, we examined Lagrangian stability of a plane parallel
class of curves defined by the resultant equation. Based on this ex
tension, we examined Lagrangian stability of a plane parallel
flow. Investigation of more complex basic flows, such as
the Arnold-Beltrami-Childress (ABC) flows
\[ U = (A \sin z + C \cos y, B \sin x + A \cos z, C \sin y + B \cos x), \]
will be the subject of future work.

### Acknowledgements

The author is grateful to Professor Tsutomu Kambe and Associate Professor Fujihiro Hamba for useful comments. The ABC flows are inviscid (not viscous) flows, but it should be noted that Claim 8 is also valid for the inviscid case.

This work was supported in part by the Special Postdoctoral Researchers Program of RIKEN (The Institute of Physical and Chemical Research).

### References


Maximum amplitude and instability of a shallow water solitary wave

S. Watanabe1, M. Oikawa2, H. Tsuji2, Y. Funakubo3, K. Ohtsuka3 and T. Minagawa1

1 Department of Physics, Yokohama National University, 79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan
2 Research Institute for Applied Mechanics, Kyushu University, 6-1 Kasuga-koen, Kasuga, Fukuoka 816-8580, Japan
3 Department of Energy Engineering, Yokohama National University, 79-5 Tokiwadai, Hodogaya, Yokohama 240-8501, Japan

The instability of a shallow water solitary wave is observed experimentally for the first time. The instability takes place when the amplitude and the velocity exceed the normalized maximum values, 0.70 and 1.29, for an observed stable solitary wave. Due to the instability, ripples are excited from the crest, while the solitary wave amplitude decreases and a plateau behind a solitary wave is formed gradually. The breaking of ripples of micro scale is observed in the leading edge. The results are compared with the previous theories.

Despite a long history of studies on shallow water solitary waves, not many experiments aiming at finding a shallow water solitary wave of large amplitude have been reported. Daily and Stephans [1] succeeded to observe a solitary wave with amplitude of \(a/h=0.62\), where \(a\) and \(h\) denote the wave height and the water depth. Renanard, Sebra Santos and Temperville [2] generated the wave with initial amplitude of \(a/h = 0.65\), which is the maximum solitary wave ever reported. The significance of finding a shallow water solitary wave of large amplitude lies in verifying the theory predicting the instability of a solitary wave.

Theoretical efforts to explore the nonlinear properties of shallow water waves have been devoted from various points of view since the discovery of the K-dV equation.[3] One of the important results from higher order perturbation theories by Fenton [4] and by Longuet-Higgins and Fenton [5] is that the energy, momentum and so on have the maxima when the amplitude of a solitary wave is increased. These results lead to the problem of stability of a solitary wave. Tanaka [6] showed through the method of an integral equation that a solitary wave becomes unstable in case the wave height exceeds the amplitude of the maximum energy. Tanaka and his collaborators [7] later verified the instability numerically. It is noted that these theoretical studies are based on the potential theory and neglect viscous effect, capillary effect and so on.

The maximum amplitude of a stable solitary wave by Tanaka [6] is \(a/h=0.780640\) and the velocity corresponding to this wave height is \(C_0(gh)^{1/2} = 1.29385\) where \(C_0\) and \(gh^{1/2}\) denote the velocity of a solitary wave and the linear wave velocity in the long wavelength limit in terms of the gravitational acceleration \(g\). Hereafter we call these amplitude and velocity the critical amplitude and the critical velocity.

The aim of the present paper is to observe the maximum height of a stable solitary wave propagating along a water tank of uniform depth, and also to observe the instability of solitary wave with the height beyond the maximum amplitude of a stable wave. So far, this problem was not investigated experimentally. For this end, a water tank of 17 m in length, 0.8 m in height and 0.6 m in width is employed. The region of measurement is limited to a half of the tank length, ~9 m, where the depth is uniform and no water inlet/outlet exists. The water depth is between 0.05 m and 0.1 m. To excite shallow water waves in a tank, we make use of a piston-type wave maker controlled by a computer. A motor installed on upper supports of a water tank drives a piston, or a paddle, of 0.58 m in width and 0.4 m in height. The temporal displacement \(d(t)\) of the piston is programmed in the following function,

\[
d(t) = \begin{cases} 
\frac{A}{2} \tanh(2\pi(t/t_0 - \frac{1}{2})) + \frac{A}{2} \tanh \pi, & 0 \leq t \leq t_0 \\
A \tanh \pi, & t \geq t_0 
\end{cases}
\]

where \(A\) and \(t_0\) denote the magnitude and period of displacement. The argument of hyperbolic tangent function changes from \(-\pi\) (at \(t=0\)) to \(\pi\) (at \(t=t_0\)) with the time \(t\), so that the piston moves by 0.996\(d\) where 0.996=\(\tanh \pi\). The motion of the piston is determined by two parameters, \(A\) and \(t_0\). We choose \(0.1\) (m) \(\leq A \leq 0.25\) (m) and \(1\) (sec) \(\leq t_0 \leq 2\) (sec). To detect waves, nine pieces of wave gauges of capacitance-type are placed in the intervals of 1 m from the wave maker. Waves are monitored by a digital video camera.

We have carefully chosen the combination of the magnitude and period of displacement, \(A\) and \(t_0\), for a given water depth \(h\) in order to generate a single solitary wave of desired amplitude with the suppression of oscillating wave tail. A typical example of propagation of a stable solitary wave in shallow water is shown by dotted curves in Fig. 1 where the wave height is depicted as a function of time. The observed wave is not symmetric with respect to the wave peak near the wave maker, but evolves gradually to a symmetric solitary waves after leaving oscillatory waves of small amplitude behind.

The wave height slowly decreases with the distance with a small amount, 2 or 3 %, of fluctuation that comes from the characteristics of each wave gauge. The normalized wave height, \(a/h\), is 0.66 at 9 m, which is larger than the height reported previously.[1,2] In the experiment
of Renauard et al., the initial damping of a large amplitude solitary wave is heavy, that is, the height $a/h = 0.65$ at 1 m from the wave maker reduces to $a/h = 0.5$ at 10 m. So their initial damping is about 20% in that interval, while, in our case, the damping is less than 7% in the interval of 8 m.

The velocity of a stable solitary wave is plotted as a function of wave amplitude in Fig. 2. The experimental results are obtained after symmetric waveform has been attained, that is, the velocity and wave height are observed at 8 and 9 m, and the wave height is averaged value. They agree with the theoretical curve [5] especially in a small amplitude region, but systematically deviate from it in a large amplitude region. In most cases, the velocity is larger than the theory when the wave height approaches the critical value. Such a deviation presumably comes from the capillary effect, which has been neglected theoretically.

We see from Fig. 2 that the maximum amplitude of a stable solitary wave is $a/h = 0.7$ and the maximum velocity is $c_0(gh)^{1/2} = 1.29$, or 1.285 to be exact. Both of them are smaller than the critical values for the onset of the instability as was predicted by the theory. The former, however, is considerably smaller than the critical amplitude, while the latter is close to the critical velocity.

Next, we consider the stability problem of a solitary wave. The propagation of an unstable solitary wave is shown by solid curves in Fig. 1, where the period of piston displacement is shortened by 0.1 sec compared to a stable wave in order to enhance wave height and velocity. The evolution of an unstable wave differs subtly from the stable one. 1) Although the amplitudes observed at 1 and 2 m from the wave maker are larger than a stable wave, the wave height suddenly decreases at 3–5 m and simultaneously starts to deviate from symmetrical forms. 2) Ripples excited at the peak of a solitary wave travel toward leading and trailing edges and produce the rough surface of wave. (The ripples are shown later in Fig. 4) The ripples going down the leading edge soon cease traveling near the half maximum of the solitary wave, but the ripples going down the trailing edge travel far behind a solitary wave. 3) A plateau appears at 4 m in the trailing edge of a solitary wave and spreads backward at 5–7 m. 4) At 8 m from the wave maker, however, the wave becomes smooth again, that is, no more ripple is generated from a solitary wave. The plateau gradually separates from a main solitary wave and evolves to the second but small solitary wave. 5) Low frequency audible noise is generated while ripples are excited at the peak of a solitary wave, contrary to silent propagation of a stable solitary wave. 6) The amplitude of a solitary wave at 9 m reduces to $a/h = 0.54$.

Among these properties, the first four facts indicate that the instability of a solitary wave occurs between 2 and 3 m from the wave maker, and the stability is recovered between 7 and 8 m. The instability is characterized by the generation of ripples accompanied by audible noise, the decrease of wave height and the consequent formation of plateau.

The above properties are common to the propagation of unstable solitary wave, except that the positions for the onset of instability and for the recovery of stability depend sensitively on the initial conditions. Generally the position for the onset of instability approaches to the wave maker, if the initial wave height is increased. As for the position for the recovery of stability, it is difficult to draw general conclusions, because the position is not determined only by the initial condition. So is the wave amplitude after the recovery of stability.

The spatial evolution of normalized amplitude and velocity are shown in Fig. 3, where the stable and unstable propagations correspond to Figs. 1. In the stable case, the observed wave height and velocity are always below the critical values indicated by dotted lines. Although the velocity is close to the critical velocity, the wave height is sufficiently smaller than the critical amplitude, as was expected from the observed nonlinear dispersion relation in Fig. 2. Even in a stable propagation of solitary wave, we often encounter the case where the velocity exceeds the critical velocity near the wave maker, but goes down the critical value within a short distance. The reason for the stability will be described below.

In the unstable case, on the other hand, the wave velocity exceeds the critical velocity as far as about 5 m from the wave maker, and a solitary wave becomes unstable between 2 and 3 m in spite of the fact that the wave height is always smaller than the critical amplitude, suggesting that the instability is governed by the velocity of a solitary wave, but not by the wave height. The wave height goes down quickly after the onset of instability, but keeps almost a constant value from 5 to 9. The velocity of a solitary wave reduces slowly even after the onset of instability, reaches below the critical velocity near 5 m and then reduces rather quickly until about 8 m. It should be noted that the instability sets in only after the wave travels as long as 2 m or more with the velocity larger than the critical velocity and it is over only after the wave is propagated for more than 2 m with the velocity smaller than the critical velocity. There exists a delay time for the onset and also the end of instability. This delay time explains the result described at the end of the previous paragraph that a solitary wave keeps stable even when the velocity is above the critical velocity only in the vicinity of the wave maker (in the interval less than 2 m).

The development of instability is shown in Fig. 4 (A) and (B). In the first picture, we see that a solitary wave is smooth and stable in the one third of the width of a tank near the rear sidewall, but the remainder of the wave toward the front sidewall is mixed with randomly phased ripples and is thus unstable. The instability of a solitary wave does not take place simultaneously along the wave peak, but starts from the vicinity of one of the sidewalls. In the present experiment, the instability is always initiated from the front sidewall and is propagated transversely toward the rear wall. This asymmetry may come from unexpected asymmetries of experimental setup, such as the imperfection of perpendicularity of piston of wave maker with respect to sidewall, and/or of centrality of piston between sidewalls and so on. In the region where the
ripples are generated, the wave height is reduced compared to the region without ripples, in accordance with the results by wave gauges shown in Fig. 1. It is also recognized from the boundary of stable and unstable regions that the ripples are generated from the peak of a solitary wave and spread over the leading edge. The ripples are considered to be capillary waves, because the wavelength is much smaller than the solitary wave. Although the velocity of capillary wave with small wavelength is expected to travel faster than a shallow water wave, the ripples never move ahead of the solitary wave but travel only until near the half maximum of the wave height. In Fig. 4 (B), we clearly observe that the ripples of micro scale are breaking while going down the leading edge. As for the ripples traveling down the trailing edge toward the wave maker, the amplitude decreases gradually and only the long wavelength components remain as they travel far behind the solitary wave.

Low frequency audible noises are detected only when a solitary wave is unstable and ripples are continuously excited from the peak. The noises are probably produced when the overturned fluid hits the fluid beneath itself. Otherwise no noise is expected. The ripples propagating backward do not produce any noise.

In the inlet in Fig. 1, the waveform of an unstable solitary wave has been compared with that of a stable solitary wave with nearly the same amplitude. The stable solitary wave indicated by the broken curve is symmetric with respect to the peak, while the unstable solitary wave indicated by the solid curve has a steep leading edge and a gentle trailing edge. We notice therefore that the instability is closely related to the steepening of a wave.

Now, we investigate the process of steepening in Fig. 4 (C) and (D). Near the front wall, ripples are already generated, but near the rear wall, a solitary wave is still stable and no ripple is observed. We do not notice a remarkable asymmetry in the waveform of Fig. 4 (C) although the instability generating ripples has already started. In (D), however, the asymmetry resulting from the steepening in the leading edge becomes appreciable. This process means that the wave steepening takes place after the beginning of instability that generates the ripples. It is also noted that the ripples are going to break as shown in (D). The experimental observation is summarized as follows: at first the instability accompanied by the ripples takes place at the crest of a solitary wave, then the solitary wave steepens near the crest and finally the ripples break. We have never observed the breaking of wave crest.

It is worthwhile to compare the present result with the numerical solution by Tanaka et al.[7]: They pointed out two kinds of instabilities, a breaking case and a damping case depending on the sign of an initial perturbation. In the former case, a solitary wave grows in a linear stage, eventually steepens near the crest in a nonlinear stage and finally overturns at the crest. The process of wave steepening and breaking is different from the experimental results mentioned just above. In the latter case, an unstable solitary wave smoothly evolves to a stable solitary wave of lesser amplitude but very nearly the same energy. The steepening of wave does not occur in this case. In the experiment, the smooth transition to a wave of lesser amplitude without wave steepening and instability has not been observed. Therefore our result does not correspond to any of the two cases. In spite of the discrepancy, we claim the similarity to the breaking case, because of the wave steepening and of the wave breaking. One of the origins for the discrepancy might be accounted for the scale of wave. A wave of large scale in deeper water may enable us to escape from the capillary effect and lead to realize the instability predicted by the numerical solution.

In summary, we carried out the experiment on the maximum amplitude and the instability of a shallow water solitary wave. The maximum solitary wave we have observed has the normalized amplitude of 0.70 and the corresponding normalized velocity of 1.29. The instability of a shallow water solitary wave is accompanied by the generation of ripples, the steepening of the leading edge and the breaking of ripples.

The present experiment has been carried out under the collaborating research program at the Research Institute for Applied Mechanics, Kyushu University. The authors are grateful to Prof. A. Masuda for his encouragement and also to Mr. K. Marubayashi and Mr. M. Ishibashi for their technical assistance and advice.

[1] J. W. Daily and S. C. Stephens Jr., Trans ASCE. 118 575 (1953). A part of this paper is reproduced in Fig. 7 of Ref. 5.
Fig. 2  The velocity $c_0/(gh)^{1/2}$ of a stable solitary wave as a function of the amplitude $a/h$. The water depths are 10 cm (○), 9 cm (△), 8 cm (□), 7 cm (△) and 5 cm (+). The solid curve indicates the theoretical value by Longuet-Higgins and Fenton (Ref. 5). The crossed point of two dotted lines denotes the maximum amplitude and velocity of a stable solitary wave by Tanaka (Ref. 6).

Fig. 1  The propagation of stable (dotted curve) and unstable (solid curve) solitary waves. The magnitude and period of piston displacement are $A=0.25$ m and $t_0=1.2$ sec (stable) or $t_0=1.1$ sec (unstable). The water depth $h$ is 0.1 m. The origin of the abscissa is arbitrarily shifted. In the inlet at the bottom left, the waveform of an unstable solitary wave observed at 6 m (solid curve) is compared with a stable one (dotted curve) of nearly the same amplitude.

Fig. 3  The spatial evolution of normalized amplitude $a/h$ and velocity $c_0/(gh)^{1/2}$ obtained from Fig. 1. (○): Stable wave. (△): Unstable wave. The dotted line indicates the critical amplitude or velocity.

Fig. 4  The front view, (A) and (B), and the side view, (C) and (D), of the evolution of instability. A solitary waves is propagating from right to left between 6 m and 7 m from the wave maker, which is located at the top right in (A) and (B), but is not seen. The arrows in (C) and (D) indicate the stripes of capillary waves. The magnitude and period of piston displacement are $A=0.25$ m and $t_0=1.5$ sec. The water depth $h$ is 0.08 m. The time interval between the pictures is 2/15 sec for (A) and (B) and is 1/15 sec for (C) and (D).
Nonlinear Photonic Interactions among Electric Multipoles in Quantum Dots

Hideaki Matsueda

Department of Information Science, Kochi University,
2-5-1 Akebono-cho, Kochi 780-8520, Japan
Email: matsueda@is.kochi-u.ac.jp

Abstract—A comparison among the possible nonlinear photonic interactions for coherence retention in solids and for scalable quantum gates is made theoretically, and then numerical plottings are given, on the basis of the dipole length estimated from our microphotoluminescence spectra of GaAs/AlGaAs coupled quantum dots (QDs) having a pair of 0.3meV splittings.

1. Introduction

We have been interested in expanding the freedoom of mankind to harness fast and seemingly weak correlations among quantum dots (QDs) ensemble, manifesting them at macroscopic level using the recent technological progress in nanometer structures and measurements [1, 2]. Especially, the inclusion of the resonance dynamic dipole-dipole interaction (RDDDI, where excitation energy or the number of photons sufficient only for some sites out of numbers of sites is supplied) has been our major concern both in theory and in experiment, as the method to enhance coherence as well as a mean to execute logic operations, proposing a prospecitve model of solid state integrated circuits for quantum computation [1, 2, 3].

In this paper, a theoretical outlook of the possible nonlinear interactions involving virtual [4] as well as real photons will be given first. Interactions of all the combinations of a single charges, dipoles, and quadrupoles are considered, i.e. each multipole exerts its potential on other multipoles yielding the interaction energies \( W_{cd}, W_{dd}, W_{ds}, W_{dq}, W_{qs}, W_{qd}, W_{qq} \), where the first suffix refers to Colonomic (c), dipolar (d), and quadrupolar (q) potentials respectively, and the second to a single charge (s), dipole (d), and quadrupole (q) moreover. The second order perturbation energy due to the dipole-dipole interaction \( W_{dd+q} \) is also included, since this amounts to significant energy at very close distance as van der Waals force. Then, numerical plottings are given for these interactions, on the basis of dipole length estimated by our microphotoluminescence spectra of GaAs/AlGaAs coupled QDs [2].

2. Comparison of Multipole Interactions

2.1. Derivation of the Multipoles

As represented in Fig.1, an arbitrary distribution of electric charges \( e_i \) at position \( r_i = e_x x_i + e_y y_i + e_z z_i \) imposes a potential \( \Phi_c \) at position \( R = e_x X + e_y Y + e_z Z \) following the Coulomb law, where \( e_x, e_y \) and \( e_z \) are the unit vectors of the Cartesian coordinates.

\[
\Phi_c = \sum_i \phi_i (|R - r_i|) = \sum_i \frac{e_i}{4\pi \epsilon |R - r_i|} 
\]

\[
= \sum_i \frac{e_i}{4\pi \epsilon} \left| \frac{e_x (X - x_i) + e_y (Y - y_i) + e_z (Z - z_i)}{r_i} \right| 
\]

\[
= \sum_i \frac{e_i}{4\pi \epsilon} \sqrt{(X - x_i)^2 + (Y - y_i)^2 + (Z - z_i)^2} 
\]

in SI unit. The Coulomb potential may be expanded in three dimensional Taylor series as follows, usually assuming that \( r_i \) is much smaller than \( R \), i.e. \(|r_i| \ll |R|\).

\[
\Phi_c = \sum_i \frac{e_i}{4\pi \epsilon} \frac{r_i}{|R - r_i|} \left[ x_i \frac{\partial}{\partial x_i} \bigg|_{x_i=0} + y_i \frac{\partial}{\partial y_i} \bigg|_{y_i=0} + z_i \frac{\partial}{\partial z_i} \bigg|_{z_i=0} \right] n 
\]

\[
+ \frac{1}{8\pi \epsilon} \sum_i \frac{e_i}{|R - r_i|} \left( \frac{1}{r_i} \right) \left[ \sqrt{(X - x_i)^2 + (Y - y_i)^2 + (Z - z_i)^2} \right]^3 
\]

\[
\times \left[ (x_i^2 - \frac{r_i^2}{3}) (X - x_i)^2 + y_i (X - x_i) (Y - y_i) \right]_{x_i=0, y_i=0} 
\]

\[
+ z_i (X - x_i) (Z - z_i)_{x_i=0, z_i=0} + y_i z_i (Y - y_i) (Z - z_i)_{y_i=0, z_i=0} 
\]

\[
+ \left( y_i^2 - \frac{r_i^2}{3} \right) y_i (Y - y_i)^2 + z_i x_i (Y - y_i) (Z - z_i)_{z_i=0, y_i=0} 
\]

\[
+ z_i x_i (Z - z_i) (X - x_i)_{z_i=0, x_i=0} + z_i y_i (Z - z_i) (Y - y_i)_{y_i=0, x_i=0} 
\]

\[
+ \left( z_i^2 - \frac{r_i^2}{3} \right) (Z - z_i)^2 \right] + \cdots 
\]

\[
= \frac{1}{4\pi \epsilon} \sum_i \frac{e_i}{|R - r_i|} + \frac{1}{8\pi \epsilon} \sum_i \frac{e_i r_i \cdot n_i}{|R - r_i|^2} 
\]

939
\[ \Phi, \quad E = \frac{d\Phi}{dR} \]

Figure 1: Coulomb potential at \( R \) exerted by a charge distribution \( e_i \) at \( r_i \). The \( \alpha_i \) is the angle between the vectors \( R \) and \( R - r_i \).

\[ + \frac{1}{8\pi\epsilon} \sum_i \frac{3e_i}{|R-r_i|} \times (X-x_i, Y-y_i, Z-z_i)_{x_i, y_i, z_i=0} \]

\[ \times \left( \begin{array}{ccc} x_i^2 - \frac{z_i^2}{3} & xy & xz \\ yx & y_i^2 - \frac{x_i^2}{3} & yz \\ zx & zy & z_i^2 - \frac{x_i^2}{3} \end{array} \right) (X-x_i, Y-y_i, Z-z_i)_{x_i, y_i, z_i=0} \]

\[ = \frac{1}{4\pi\epsilon} \sum_i \frac{e_i}{|R|} \cdot \nabla_{x,y,z} \phi_i(|R-r_i|)_{r_i=0} \]

\[ + \frac{3}{8\pi\epsilon} \sum_{mn} (R-r_{1m}) \cdot e_i Q_{mn} (R-r_{1n})_{|R-r_i|^6} + \cdots \]

where \( m \) and \( n \) correspond to \( x, y, z \)

\[ = \sum_i \phi_{ci}(R) + \sum_i r_i \cdot E_{eci} \]

\[ + \frac{3}{8\pi\epsilon} \sum_i \frac{e_i}{|R|} \cdot Q_{ci} \cdot (R-r_i)_{|R-r_i|^6} + \cdots \]

\[ = \sum_i \phi_{ci}(R) + \frac{1}{4\pi\epsilon} \sum_i p_i \cdot \nabla_{x,y,z} \frac{1}{|R-r_i|} \]

\[ + \frac{3}{8\pi\epsilon} \sum_i \frac{e_i}{|R|} \cdot Q_{ci} \cdot n_i_{1} \cdot n_i_{2} + \cdots \]

where the terms of \( \frac{e_i}{|R|} \), as well as \( x_i^2, y_i^2, z_i^2 \), and \( p_i = e_i f_i \) (in sans-serif) are not operated by the differentiation with the position coordinates such as \( x_i, y_i \) and \( z_i \) in \( \nabla \), because these have only the role of increments in the Taylor expansion, however the absolute value of \( r_i \) i.e. \( |r_i| \) refers to the dipole length \( r_d \); and \( n_i \) is the unit vector along \( R - r_i \). \( E_{eci} \) is the electric field from a simple charge \( e_i \), i.e. \( E_{eci} = \frac{e_i}{4\pi\epsilon|R-r_i|^2} \), and \( e_i Q_i \) is the quadrupole moment tensor which consists of \( Q_{i\alpha\beta\gamma} \), where \( n_i = (h,k,l) \) refers to \( R - r_i = e_x (X-x) + e_y (Y-y) + e_z (Z-z) = (X-x, Y-y, Z-z) \) and \( [R-r_i] = (h,k,l) \) with the direction cosine \( h, k, l \), meaning \( n_{i\alpha} = (h,0,0) \) or \( (0,k,0) \) or \( (0,0,l) \).

Alternatively from eqn.(4) the potential may be expanded in terms of the Legendre polynomials \[ P(|R|) \] where \( \cos \theta_i = \frac{R^2+r_i^2-|R-r_i|^2}{2r_i R} \), with \( R = \sqrt{x^2+y^2+z^2} \) and \( r_i = \sqrt{x_i^2+y_i^2+z_i^2} \), as is seen from Fig.1.

2.2. Multipole Potentials

As usual, the major parts of the first three terms of eqn.(7) or (8) or (9) are assigned to the simple Coulombic potential \( \phi_{ci} \), the dipole potential \( \phi_{di} \), and the quadrupole potential \( \phi_{qi} \) respectively at \( R \), originated from a single charge at position \( r_i \) as follows.

\[ \phi_{ci} = \frac{e_i}{4\pi\epsilon} \frac{1}{|R-r_i|} \Rightarrow \frac{e_i}{4\pi\epsilon} \frac{1}{R} \]

\[ \phi_{di} = \frac{1}{4\pi\epsilon} \frac{p_i \cdot n_i}{|R-r_i| |r_i|^2} \Rightarrow \frac{e_i}{4\pi\epsilon} \frac{r_d \cos \theta_i}{R^2} \]

\[ \phi_{qi} = \frac{3}{8\pi\epsilon} \frac{(R-r_i) \cdot e_i Q_{i} \cdot (R-r_i)}{|R-r_i|^6} \]

\[ = \frac{3}{8\pi\epsilon} \frac{n_i \cdot e_i Q_{i} \cdot n_i}{|R-r_i|^6} \]

\[ = \frac{e_i}{8\pi\epsilon} \frac{r_i^2 (3 \cos^2 \theta_i - 1)}{R^3} \]

where the quadrupole moment is \( e_i Q_{i} \), and the arrow \((\Rightarrow)\) corresponds to the asymptotic relation \(|R-r_i| \Rightarrow R\).

2.3. Interactions of Multipoles with Multipole Potentials

These multipolar potentials in turn interact with other multipolar potentials \( e_j, \) a dipole (vector) \( p_i = e_j f_j \), and a quadrupole \( Q_j \), yielding the interaction energies \( W_{ci}, W_{ci}, W_{ci}, W_{di}, W_{di}, W_{di}, W_{qi}, W_{qi}, W_{qi}, \) and \( W_{qi} \), where cs, cd, qa, ds, dq, qs, dq, and qd refer to the combination of the potentials and the charge, dipole, or quadrupole; e.g. \( W_{cs} \) represents the interaction of the simple Coulombic potential \( \phi_{ci} \) with a single charge \( e_j \), \( W_{dd} \) the dipole potential \( \phi_{di} \) with another dipole \( p_j = e_j f_j \), and so on. The detailed expressions of these are given below in eV and SI units, and also plotted as functions of the inter-polar distance \(|R-r_i| \equiv |R| = R\) between the multipoles in Fig.1, on the basis of the dipole length \( r_d = |r_i| = 1.397nm \) estimated by
our microphotoluminescence spectra of GaAs/AlGaAs coupled QDs showing a pair of 0.3meV splittings [2].

\begin{equation}
W_{cs} = \phi_{c^1} \cdot \phi_{c^2} \cdot \frac{1}{e} \Rightarrow \frac{e_i}{4\pi\epsilon R}
\end{equation}

\begin{equation}
W_{cd} = \nabla \phi_{c^1} \cdot \phi_{d^1} \cdot \frac{1}{e} \Rightarrow \frac{e_i r_d \cos \beta}{4\pi\epsilon R^2}
\end{equation}

\begin{equation}
W_{cq} = \left[ \nabla^2 \phi_{c^1} \cdot \phi_{c^2} \cdot \frac{1}{e} \right] \Rightarrow 0 \text{ for } |r| << |R|
\end{equation}

Multipolar interaction energies of all the nontrivial combinations $W_{cs}, W_{cd}, W_{ds}, W_{dd}, W_{dq}, W_{qs}, W_{qd}, W_{qq}$, in eqn.(19) may be estimated, using this equation as that perpendicular to the line.

\begin{equation}
W_{dd}\| = \left[ \nabla \phi_{d^1} \cdot \phi_{d^2} \cdot \frac{1}{e} \Rightarrow \frac{e_i r_d^2}{2\pi\epsilon R^3} \right]
\end{equation}

\begin{equation}
W_{dd\perp} = \left[ \nabla \phi_{d^1} \cdot \phi_{d^2} \cdot \frac{1}{e} \Rightarrow -\frac{e_i r_d^2}{4\pi\epsilon R^3} \right]
\end{equation}

\begin{equation}
W_{dq} = \left[ \nabla^2 \phi_{d^1} \cdot \phi_{q^1} \cdot \frac{1}{e} \right] \Rightarrow -\frac{3e_i r_i r_d n_i \cdot n_j}{2\pi\epsilon R^3} Q_j
\end{equation}

\begin{equation}
W_{qs} = \phi_{q^1} \cdot \phi_{c^2} \cdot \frac{1}{e} \Rightarrow \frac{3e_i}{8\pi\epsilon} n_i \cdot \frac{n_j}{R^3} Q_i
\end{equation}

\begin{equation}
W_{qd} = \nabla \phi_{q^1} \cdot \phi_{q^2} \cdot \frac{1}{e} \Rightarrow \frac{3e_i}{8\pi\epsilon} r_i \cdot \frac{Q_i Q_j}{R^4} r_d \cos \beta
\end{equation}

\begin{equation}
W_{qq} = \left[ \nabla^2 \phi_{q^1} \cdot \phi_{q^2} \cdot \frac{1}{e} \right] \Rightarrow \frac{15e_i}{2\pi\epsilon} n_i \cdot \frac{n_j}{R^3} Q_i
\end{equation}

where $dd\|$ and $dd\perp$ correspond respectively to the dipoles parallel to the line connecting the two, and that perpendicular to the line.


As the natural consequence, the second order perturbation effect of the dipole-dipole interaction given in eqn.(19) may be estimated, using this equation as the perturbing Hamiltonian $\mathcal{H}_{dd}(i, j)'$. According to the standard perturbation theory [6], it may be possible to introduce a parameter $\lambda$ which is smaller than 1 ($\lambda \ll 1$) representing the strength of the perturbation, in front of the perturbing Hamiltonian $\mathcal{H}_{dd}(i, j)'$ comparable to the unperturbed Hamiltonian $\mathcal{H}_0$ giving $\mathcal{H}_{dd}(i, j) = \lambda \mathcal{H}_{dd}(i, j)'$, and to express the new eigenstates (eigenvectors) $|\Psi_\pm\rangle$ and the eigenvalues $E_\pm$ in series of $\lambda$, using a tentative set of basis $|\Psi_0\rangle$ and $|\Psi_0'\rangle$.

\begin{equation}
\mathcal{H} = \mathcal{H}_0 + \lambda \mathcal{H}_{dd}(i, j) = \mathcal{H}_0 + \lambda \mathcal{H}_{dd}(i, j)'
\end{equation}

\begin{equation}
\mathcal{H} (|\Psi_+\rangle) = \mathcal{H}_0 (|\Psi_\pm\rangle)
\end{equation}

\begin{equation}
|\Psi_+\rangle = |\Psi_0\rangle + \lambda^2 |\Psi_0'\rangle + \lambda \Psi_0'' + \cdots
\end{equation}

\begin{equation}
E_+ = E_0 + \lambda^2 E_0' + \cdots
\end{equation}

assuming the convergence of these power series in $\lambda$.

The original eigenstates of unperturbed Hamiltonian $\mathcal{H}_0 = H_{0i} + H_{0j}$ are represented as $|\Psi_0\rangle$ and $|\Psi_0'\rangle$ [2].

If we put these expansion of $|\Psi_\pm\rangle$ and $E_\pm$ into the time-independent Schrödinger equation eqn.(25), equations polinomial in $\lambda$ that should be satisfied by small but all possible values of $\lambda$ are derived. Then comparing the terms of same order in $\lambda$, an equation of the second order perturbation energy $W_{dd\perp p}$ is derived.

\begin{equation}
W_{dd\perp p} = \frac{W_{dd\perp}}{\Delta} \Rightarrow \frac{\epsilon_i^2 r_d^2}{4\pi\epsilon R^3} \frac{\Delta}{16\pi^2 \epsilon^2 \Delta R^6}
\end{equation}

where $\Delta$ is the detuning of the original unperturbed levels, i.e. $\Delta = E_r - E_l$ in our coupled quantum dots.

4. Numerical Demonstration of Multipolar Interactions based on the Experimentally Estimated Dipole Length

The maximum energies of the different multipolar interactions in eV as the functions of the inter-polar distance in nm are calculated (see Fig.2) on the basis of the dipole length in the quantum dot $r_d = |r| = r_i$ (or $r_j$) = 1.39nm, estimated by our microphotoluminescence spectra of GaAs/AlGaAs coupled QDs having 0.3meV splittings [2], which is about four times larger than the dipole length of bulk GaAs [7]. It is understood that $R_l = |R - r_i|$, and $\epsilon = \epsilon_{GaAs} - \epsilon_{clad} = 12.9\epsilon_o$, $\epsilon = 10^7 \times 8.54 \times 10^{-12} C^2 N^{-1} m^{-2}$. $e = 1.602 \times 10^{-19} C = 4.803 \times 10^{-10} esu$. The dipole moment is estimated as $e r_d = (1.602 \times 10^{-19} \times 1.397 \times 10^{-9} C \times m = 6.706 \times 10^{-17} esu \times cm = 67.06 Debye$. The displacement of the quadrupole is assumed to be 0.3 of the dipole length, i.e. $r_{q-pol} = 0.3 r_d$, then $e Q = e r_d^2 q_{pol} = 2.814 \times 10^{-38} \times 8.431 \times 10^{-25} esu \times cm^2$. The second order perturbation energy $W_{dd\perp p}$ due to the dipole-dipole interaction between the dipoles perpendicular to the line of connection is also shown for the detuning of the original unperturbed levels $\Delta = E_r - E_l = 40 meV$ in our coupled quantum dots.

4.5. Concluding Remarks

A theoretical outlook of the nonlinear photonic interactions possibly useful for coherence retention in solids and for scalable quantum gates are given. Multipolar interaction energies of all the nontrivial combinations $W_{cs}, W_{cd}, W_{ds}, W_{dd}, W_{dq}, W_{qs}, W_{qd}, W_{qq},$
Figure 2: Absolute values of the multipolar interactions in eV, as functions of the inter-polar distance, \( R_i = |\mathbf{R}_i - \mathbf{R}_j| \). The dipole length \( r_d = r_d \) (or \( r_j = d_{GaAs}QD = 1.397\text{nm} \)).

and the second order perturbation energy due to the dipole-dipole interaction \( W_{dd,\perp} \) are demonstrated as functions of the inter-polar distance, using our experimental estimation of the dipole length [2].

It is seen from Fig.2 that below 6 ~ 8 nm distance the energy of dipole-dipole interaction start to exceed 1 meV, and the interaction becomes resolvable by conventional spectroscopy. Below 3 nm the second order perturbation energy of dipole-dipole (van der Waals) interaction start to exceed 1 meV. Furthermore, real or virtual photons may be assigned as the mediator for the multipolar interactions [4], giving a possibility of novel controlling technology in nanometer structures.

Acknowledgments

The author is grateful to Prof. Jorn M. Hvam and his group at Technical University of Denmark especially for the microphotoluminescence measurements, and to Prof. Eli Kapon and his group at Swiss Federal Institute of Technology Lausanne especially for the fabrication of coupled quantum dots.

References


Abstract—We present a fixed word-length adaption of the ternary search tree algorithm originally proposed by Sedgewick and Bentley that allows the fast computation of the boxcounting statistics in arbitrary high dimensions.

1. Introduction

Boxcounting is a multivariate histogram technique where a high-dimensional space is divided into rectangular partitions (boxes). The number of data set points falling into each partition is counted and can serve as a histogram-like approximation to the underlying distribution. By varying the size of the partitions the scaling behavior of the number of non-empty boxes is used to estimate the boxcounting dimension. The boxcounting dimension (also called capacity dimension) is an important quantity for characterizing fractals and chaotic attractors [1]. Furthermore, several theorems justifying state space reconstruction using delay coordinates contain bounds of relevant embedding dimensions referring to the boxcounting dimension of the underlying attractors [2]. Due to this theoretical importance efficient algorithms for estimating the boxcounting dimension in high dimensions are required, because naive implementations result in exorbitant computational costs that often prevent applications of this fractal measure. Therefore, we introduce here the ternary search tree algorithm as proposed by Sedgewick and Bentley [3] as an efficient technique for histogram based methods in multivariate statistics. A ternary search tree (TST) is a dictionary data structure specially designed for situations where the key is a string. A TST allows to efficiently locate, insert and delete keys. TSTs store key-value pairs, where the value can be an arbitrary object. We adapted the TST algorithm to accommodate for a fixed key length since all points of a multivariate data set share usually the same dimensionality $D$. The object associated to a key is in our adaption an integer value counting the number of points falling into the corresponding partition.

1.1. Boxcounting

A naive implementation of the boxcounting technique which keeps track of all possible partitions by means of an array in main memory quickly fails since the number of possible partitions grows exponentially fast with the data set dimension $D$.

TSTs can be seen as a multivariate extension of the binary search tree. Binary search trees allow efficient searches for scalar values. TSTs extend the binary search tree by adding an additional third pointer into the data structure coding for a tree node:

- **level**: Tree level which corresponds to the dimension ($\in \{1, \ldots, D\}$) of the key value
- **key**: Current key value. The key value is derived by partitioning ($\in \{1, \ldots, b\}$) the $\text{level}$-th coordinate of the point that is currently inserted.
- **lokid**: Pointer to the tree node at the current tree level with a smaller key value than this node.
- **hikid**: Pointer to the tree node at the current tree level with a higher key value than this node.
- **eqkid**: Pointer to the root node of the next level’s subtree. Key values of these subtree are derived from dimension $\text{level}+1$.
- **count**: Number of data set points represented by this node.

Below is a code fragment detailing the data structure in C++ notation:

```cpp
struct Tnode {
    long splitkey;
    int level;
    // tree level
    unsigned long count;
    // count how often the key leading
    // to this node exists
    Tnode *lokid;
    Tnode *eqkid;
    Tnode *hikid;
};
```

The key values for first binary tree of the TST are taken from the point’s first coordinate level = 1. Besides the pointer to nodes with higher resp. lower key values of coordinate level, the data structure keeps track of a pointer to...
the root of a subtree which sorts the key values taken from
the coordinate level + 1 of the point set (see Fig. 1).

Each partition of the space is represented by a node at
level = D. Nodes at level = D do not store a valid eqkid-
pointer, however their count field contains the number of
data set points falling into the box specified by the path
leading to this node (see Fig. 1). The total number of nodes
can not exceed the product ND of the total number of data
set points N and the dimension D. To locate a partition in
the ternary search tree given its coordinates, it is sufficient
to perform D binary searches, starting from the first binary
tree which stores key values for level = 1.

Below is a code fragment detailing the complete class
definition for a TST:

```cpp
#define TST_BUFFERS 128 // maximum number of buffers
#define TST_BUFSIZE 512 // size of first buffer
class ternary_search_tree {
protected:
  typedef Tnode *Tptr;
  // these four variables are used to
  // accelerate allocation
  // of Tnode objects
  Tptr buf;
  // pointer to current buffer
  long last_buf_size;
  // size of the buffer that was
  // recently allocated
  long next_buf_size;
  // size of the buffer that
  // will be allocated next
  long bufn;
  // number of buffers in use
  long frcn;
  // number of free nodes in
  // current buffer
  Tptr frcarr[TST_BUFFERS];

  const long len; // key length
  Tptr root; // tree root node
public:
  ternary_search_tree(const long keylength)
  : bufn(0), frcn(0), root(0),
    last_buf_size(0), len(keylength),
    next_buf_size(TST_BUFSIZE) {}
  ~ternary_search_tree() {}  // destructor
  for (long i = 0; i < bufn; i++)
    delete []frcarr[i];
  }
  int insert(const long *const key);
  // insert key vector
  void traverse(const long *count);
  // count must be of length len
  // and initialized to zero
};
```

Figure 1: Example ternary search tree into which the points
{(5, 3, 2), (1, 4, 1), (5, 3, 2), (2, 2, 3), (5, 3, 5), (1, 4, 1),
(2, 2, 3), (2, 2, 1), (2, 2, 3)} were inserted in the given order.

1.2. Node insertion and memory allocation

Memory for new nodes is allocated during the insertion
of a node. To keep overall number of calls to malloc low,
memory is allocated in chunks. Chunk sizes \( M_i \) are
increased in a Fibonacci scheme, resulting in an approxim-
ately exponential growth of the chunk sizes:

\[
M_{i+1} = M_i + M_{i-1} \quad M_0 = 0, M_1 = 512
\]

Below is a code fragment detailing the insertion of a new
point into the ternary search tree:

```cpp
int ternary_search_tree::insert(const long *const key) {
  long d;
  Tptr pp;
  Tptr *p = &root;
  long level = 0; // level goes up to len-1
  while (pp = *p) {
    // as long as we encounter already
    // existing nodes, we stay
    // inside this while loop
    if ((d = key[level] - pp->splitkey) == 0) {  // go to next tree level
      pp->count++;  // p = &pp->eqkid);
      if (++level == len)
        return 0;
    } else if (d < 0) {
      p = &pp->eqkid;
      // move left in the current level
    } else {
      p = &pp->lokid;
      // move in the current level
    }
  }
  return 0;
}
```
1.3. Linear tree traversal

Since for counting the number of occupied boxes it is not necessary to traverse the tree in a particular order, we simply scan the allocated memory blocks sequentially, thus avoiding the overhead of keeping track of the nodes that still have to be visited like commonly encountered in tree traversal. The following code fragment illustrates the computation of the boxcounting statistics for all data sets of truncated (1,...,D−1) or full dimensionality D within a single scan:

```c
void ternary_search_tree::traverse(long *ct) {
    long buf_size = TST_BUFFERS;
    // the actual size of the buffer
    // is increased each iteration
    long last_size = 0;
    long remember_size = 0;
    // traverse all buffers
    for (long i = 0; i < bufn; i++) {
        // last buffer might not be
        // filled completely
        ct[0] += count[i] =
            (i == bufn-1) ? buf_size - free_n : buf_size;
        if (free_n == 0) {
            if (bufn == TST_BUFFERS) return FAILURE;
            if (offset == 0) {
                offset = new Tnode[next_buf_size];
                freearr[bufn++] = buf;
                free_n = next_buf_size - 1;
                const long remember_size = next_buf_size;
                next_buf_size += last_buf_size;
                // Fibonacci type buffer size increase
                // this keeps overall number of
                // buffer allocations small
                last_buf_size = remember_size;
            }
            *p = buf++;
            pp = *p;
            pp->splitkey = key[level];
            pp->count = 1;
            // this node is newly created, so
            // count is set to one
            pp->level = level;
            pp->lokid = 0;
            pp->eqkid = 0;
            pp->hikid = 0;
            if (((++level) == len) return 0;
            p = &(pp->eqkid);
        }
    }
}
```

2. Performance Results

2.1. Time and space complexity

Space complexity is governed by the number of tree nodes allocated. The maximum number of nodes is limited by ND where N is the number of points in the data set and D is the data set dimension. The time complexity depends on the balancing of all subtrees within the ternary search tree. The complexity of locating a terminal node in a balanced tree is of order O(Dlog(P)), where P is the number of partitions per coordinate. Thus the time complexity of constructing the TST for a data set of N points is of order O(NDlog(P)). Unfortunately there is no guarantee that the subtrees remain balanced when points are inserted randomly into the tree. In the worst case the time complexity of a single search can deteriorate from logarithmic to linear O(DP) when subtrees are degenerated. However, techniques similar to the balancing of quick sort could also remedy the worst case behavior of the ternary search tree algorithm.

2.2. Numerical benchmarks

Numerical benchmarks were performed on a standard PC with Pentium 4 (Prescott, Stepping E0) 3.2 GHz and 1 GB of RAM running the operating system Windows XP. The TST code consists of a compiled executable for the Matlab numerical programming environment. We generated data sets of different sizes by two means:

- Numerical integration of the generalized Chua 5-

---

The code used for this experiments is available at http://zti.if.uj.edu.pl/~merkwirth/fracdims.htm and as part of the OpenTSTool package http://www.physik3.gwdg.de/tstool/ for nonlinear time series analysis.
Scroll system of differential equations:
\[
\begin{align*}
\dot{x}_1 &= \alpha(x_2 - h(x_1)) \\
\dot{x}_2 &= x_1 - x_2 + x_3 \\
\dot{x}_3 &= \beta x_2 \\
h(y) &= y m_5 + 0.5 \sum_{i=1}^{5} (m_{i-1} - m_i)(|y + c_{i-1}| - |y - c_{i-1}|)
\end{align*}
\]
with parameters \(\alpha = 9\), \(\beta = 14.286\), \(m_0 = -\frac{0.9}{2}\), \(m_1 = -\frac{1}{2}\), \(m_2 = \frac{3 \sqrt{2}}{2}\), \(m_3 = -\frac{3 \sqrt{2}}{2}\), \(m_4 = \frac{3}{2}\), \(m_5 = -\frac{3 \sqrt{2}}{2}\), \(c_0 = 1\), \(c_1 = 2.15\), \(c_2 = 3.6\), \(c_3 = 6.2\), \(c_4 = 9\). The integration starts with random initial conditions and is sampled every \(\Delta T = 0.5\). The first 20000 samples are discarded to minimize the influence of the random initial conditions. We then generated a time-delay embedding [2] of \(x_1\) with delay 1 and dimension \(D\) varying from 1 to 40.

- By means of Matlab’s `randn` we generated data sets of dimension \(D\) ranging from 1 to 40 consisting of normally distributed random numbers with zero mean and unit variance.

Both data sets were scaled and discretized to integer values in the range 0, 1, ..., 63. The computation time for scaling and discretization is included in the timing measurements below.

Figures 3 and 2 detail the dependence of the average computation time with respect to the dimension of the data set used. A linear regression analysis of the product \(ND\) versus the benchmarked computation time shows as correlation coefficient \(r^2\) of 0.992 for the Chua 5-Scroll data set and a correlation coefficient \(r^2\) of 0.999 for the random data set, indicating that the product \(ND\) which determines the space (memory) complexity seems to be dominant for everyday data sets. The Chua 5-Scroll exhibits an apparently overlinear behavior in \(D\) which is however dominated by the respective computations times for the random data set.

3. Conclusion

An algorithm is presented that efficiently solves the box-counting problem in arbitrary high dimensions, avoiding the problem of exponentially increasing space complexity of a naive implementation. An attractive feature of the algorithm is the mainly linear time complexity in the number of data set points and the data set dimension.

References


A Method of Proving Existence of Solution Curve for Nonlinear Equation using Affine Arithmetic

Yuchi Kanzawa†, Shin’ichi Oishi‡

†Department of Communication Engineering, Faculty of Engineering, Shibaura Institute of Technology
‡Department of Computer Science, School of Science and Engineering, Waseda University

Abstract—In this paper, a new algorithm is proposed to guarantee the existence of the solution curve of nonlinear systems of equations. This algorithm uses the transformation by which an axis and the tangent vector of the solution curve has the same direction. This algorithm also uses Affine Arithmetic. Affine Forna, which is the expression used in Affine Arithmetic, can express hyper-rectangles whose boundaries are not parallel to every axis. By these uses, the proposed algorithm is expected to guarantee the existence of the solution curve more effectively than conventional methods using Interval Arithmetic.

1. Introduction

Solutions of nonlinear systems of equations

\[ f(x) = 0, \quad x \in f : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n \]  

(1)

are one dimensional manifolds, called “solution curve”. Proving the existence of solution curve for Eq.(1) is the essence of tracing solution curve, which is an important problem. For example, in order to obtain characteristic curves of nonlinear resistive circuits or in order to obtain DC solutions by homotopy methods, it is necessary to solve an equation of the form (1). Homotopy methods have been studied for obtaining global convergence, and many efficient algorithm have been developed to trace solution curves [1]. Some of these algorithms can be easily performed with very little computational effort. As these algorithm also output only discrete points on the solution curve, the region cannot be obtained where the solution curve is guaranteed to exist. Consequently, they cannot avoid the danger of jumping onto different curves.

In [2], an algorithm has been proposed to obtain the region where the solution curve is guaranteed to exist and to avoid jumping onto different curves. But the algorithm can be performed only with the assumption that Lipschitz constant of \( f' \) is calculatable. The assumption is considered to restrict applied equations much.

In [3], another algorithm has been proposed to obtain the region where the solution curve is guaranteed to exist and to avoid jumping onto different curves. The algorithm can be performed without the above assumption by Interval Arithmetic. But the algorithm is considered to be not so effective because the region of the solution curve is described by interval, that is, hyper-rectangle whose each boundary is parallel to the corresponding axis. If the tangent vector \( v \) of the solution curve and one of boundaries for the region have the same direction, this algorithm is expected to be more effective.

In this paper, we propose a new algorithm to prove the existence of solution curve using Affine Arithmetic [5]. In the algorithm, the region where the solution curve is guaranteed to exist is described as hyper-rectangle whose one boundary and the tangent vector of the solution curve have the same direction, so the new algorithm is expected to me more effective than the conventional method.

In the first section, notations, definitions and the conventional method are described. In the second section, the new algorithm is proposed based on Affine Arithmetic and a linear transformation.

2. Preliminaries

In this section, some notations and definitions are introduced. In the first subsection, Interval Arithmetic is introduced and an algorithm is introduced to prove the existence of solution curve using Interval Arithmetic [3]. In the second subsection, Affine Arithmetic [5] is introduced and an algorithm is introduced to prove the existence of solution (not curve but point) of nonlinear equations [7].

2.1. Interval Arithmetic and Conventional Method

In this subsection, Interval Arithmetic is introduced and an algorithm is introduced to prove the existence of solution curve using Interval Arithmetic. Notations of interval analysis used in this paper are by Ref. [4]. Let \( I(U) \) be a set of all intervals belongs to \( U \subset \mathbb{R}^{n+1} \). Let \( \text{Rad}(I) \) be the radius of interval \( I \) (component-wise). Let \( \text{Mid}(I) \) be a midpoint of interval \( I \) (component-wise). Let \( |I| \) be the absolute value of interval \( I \) (component-wise). Let us consider the norm of interval vector \( I \) defined as

\[ \| I \|_\infty = \max_i |I_i|. \]

Let us consider the norm of \( n \times n \) interval matrix \( A \) defined as

\[ \| A \|_\infty = \| A(1, \cdots, 1) \|_\infty. \]

The following algorithm proves the existence of solution curve using Interval Arithmetic.

Algorithm 1 [3]

Step 1. Determine \( \overline{s} > 0. \)
Step 2. Obtain $v \in \mathbb{R}^{n+1}$ such that
\[
\begin{align*}
  &f'(c)v = 0, \\
  &\langle v, v \rangle = 1
\end{align*}
\]
Step 3. Apply Krawczyk method to the equation
\[
F(x, s) = \begin{cases} f(x) = 0, \\
(x - c, v) - s = 0
\end{cases}
\]
and to an adequate interval, that is,
Step 3-1. Calculate
\[
F(c, [-\overline{s}, \overline{s}]) = \left( \begin{array}{c} f'(c) \\ \begin{bmatrix} -\overline{s} & \overline{s} \end{bmatrix} \end{array} \right).
\]
Step 3-2. Choose an invertible matrix
\[
R \simeq F^{-1}(c, 0) = \left( \begin{array}{c} f'(c) \\ v \end{array} \right)^{-1}.
\]
Step 3-3. Calculate an interval
\[
T = c + 2\|RF(c, [-\overline{s}, \overline{s}])\|_{\infty}B^{n+1},
\]
where $B^{n+1}$ is the $(n+1)$-dimensional unit ball.
Step 3-4. Calculate
\[
F'(T, [-\overline{s}, \overline{s}]) = \left( \begin{array}{c} f'(T) \\ \begin{bmatrix} -\overline{s} & \overline{s} \end{bmatrix} \end{array} \right).
\]
Step 3-5. If
\[
c - RF(c, [-\overline{s}, \overline{s}]) + (E - RF'(T, [-\overline{s}, \overline{s}]))(T - c) \subseteq T
\]
holds, stop this algorithm. Then, there exists a solution curve of Eq.(1) in $T$.

If $0$ is the regular value of $f$, there exists a positive value $\overline{s}$ such that the test (2) succeeds.

Obviously each boundary of interval $T$ is parallel to the corresponding axis, while it is rare that the tangent vector $v$ of the solution curve and one of axis has the same direction. It is one of the reason why the test (2) fails for some $\overline{s}$. If the tangent vector $v$ of the solution curve and one of boundaries for test region $T$ have the same direction, it is expected that the test (2) succeed for larger $\overline{s}$.

2.2. Affine Arithmetic and Conventional Method

In this subsection, Affine Arithmetic [5] is introduced and an algorithm is introduced to prove the existence of solution (not curve but point) of nonlinear equations [7].

Let be $a_i \in \mathbb{R}$ for $i \in \{0, 1, \cdots, m\}$. Let be $-1 \leq \varepsilon_i \leq 1$ for $i \in \{1, \cdots, m\}$. The form
\[
a_0 + \sum_{i=1}^{m} a_i \varepsilon_i
\]
is called Affine Form and it describes the set
\[
\left\{ a_0 + \sum_{i=1}^{m} a_i \varepsilon_i \mid -1 \leq \varepsilon_i \leq 1, i \in \{0, 1, \cdots, m\} \right\}.
\]
Vectors whose elements are Affine Forms are called Affine Form vector. $n$-dimensional Affine Form vector $(a_0^{(1)} + \sum_{i=1}^{m} a_i^{(1)} \varepsilon_i, a_0^{(n)} + \sum_{i=1}^{m} a_i^{(n)} \varepsilon_i)$ are also described as $\alpha + A \varepsilon^{(m)}$, where
\[
\begin{align*}
\alpha &= (a_0^{(1)}, \cdots, a_0^{(n)}), \\
A &= \begin{pmatrix} a_1^{(1)} & \cdots & a_m^{(1)} \\ \vdots & \ddots & \vdots \\ a_1^{(n)} & \cdots & a_m^{(n)} \end{pmatrix}, \\
\varepsilon^{(m)} &= (\varepsilon_1, \cdots, \varepsilon_m).
\end{align*}
\]
A set of Affine Forms which describe subsets of a space $U$ is denoted by $\mathcal{A}(U)$. Absolute value of Affine Form $a_0 + \sum_{i=1}^{m} a_i \varepsilon_i$ is defined as $\max_{1 \leq i \leq n} |a_i|$. Maximum norm of $n$-dimensional Affine Form vector $a$ is defined as $\|a\|_{\infty} = \max_{1 \leq i \leq n} |a_i|$.

For $a^{(1)}, a^{(2)} \in \mathcal{A}(\mathbb{R})$, operations
\[
\left\{ a^{(1)} \ast a^{(2)} \mid * \in \{+, -, \times, \div\} \right\}
\]
and
\[
\left\{ \phi(a^{(1)}) \mid \phi \in \{\sin, \cos, \tan, \exp, \log, \cdots\} \right\}
\]
are determined as
- the result is also an Affine Form,
- the set described by the result holds,

that is,
\[
a^{(1)} \ast a^{(2)} \supset \{ x^{(1)} \ast x^{(2)} \mid x^{(1)} \in a^{(1)}, x^{(2)} \in a^{(2)} \},
\]
\[
\phi(a^{(1)}) \supset \{ \phi(x^{(1)}) \mid x^{(1)} \in a^{(1)} \}.
\]
Let $a^{(1)}$ and $a^{(2)}$ be
\[
a^{(1)} = a_0^{(1)} + \sum_{i=1}^{m} a_i^{(1)} \varepsilon_i \quad \text{and} \quad a^{(2)} = a_0^{(2)} + \sum_{i=2}^{m} a_i^{(2)} \varepsilon_i.
\]
Addition and Subtraction between $a^{(1)}$ and $a^{(2)}$ are operated as
\[
a^{(1)} \pm a^{(2)} = a_0^{(1)} \pm a_0^{(2)} + \sum_{i=1}^{m} (a_i^{(1)} \pm a_i^{(2)}) \varepsilon_i.
\]
Addition and Subtraction between $a^{(1)}$ and a constant number $c \in \mathbb{R}$ are operated as
\[
a^{(1)} \pm c = a_0^{(1)} \pm c + \sum_{i=1}^{m} a_i^{(1)} \varepsilon_i.
\]
Multiplication between \( a^{(1)} \) and \( a^{(2)} \) is operated as
\[
a^{(1)} \times a^{(2)} = a^{(1)}_0 a^{(2)}_0 + \sum_{i=1}^{m} (a^{(1)}_i a^{(2)}_i + a^{(1)}_i a^{(2)}_i) \varepsilon_i + \sum_{m=1}^{i} (a^{(1)}_i) \cdot \sum_{i=1}^{m} (a^{(1)}_i) \varepsilon_{m+1},
\]
where \( \varepsilon_{m+1} \in \mathbb{R} \) is a new symbol and satisfies \(-1 \leq \varepsilon \leq 1\). More refined variations of multiplication has been proposed [6]. Various unary operations for \( a^{(1)} \), for example, the reciprocal of \( a^{(1)} \), the square root of \( a^{(1)} \), the sine function of \( a^{(1)} \), and so on, have also been proposed but is not described here because of the sake of papers. The following algorithm proves the existence of nonlinear solution using Affine Arithmetic with related to All Solution Algorithm.

**Algorithm 2 [7]**

1. **Step 1.** Given the region of Affine Form
   \[ T = c + A \varepsilon^{(n)} \]
   where the solution is expected to exist, where \( c \in \mathbb{R}^n \) and \( A \) is a \( n \)-dimensional diagonal matrix.

2. **Step 2.** Evaluate \( f(T) \) as the Affine Form
   \[ f(T) = a + L \varepsilon + \hat{L} \varepsilon \]
   using Affine Arithmetic, where \( a \in \mathbb{R}^n \), \( L \in \mathbb{R}^{n \times n} \)
   and \( \hat{L} \) is some matrix whose number of rows is \( n \).

3. **Step 3.** Obtain \( R = L^{-1} \).

4. **Step 4.** If
   \[ -R(a + \hat{L} \varepsilon) \subset B^{(n)}, \]
   holds, there exists a solution in \( T \), where \( B^{(n)} \) is the \( n \)-dimensional unit ball.

**3. Inclusion of Solution Curve using Affine Arithmetic**

In this section, a new algorithm is proposed to prove the existence of solution curve using Affine Arithmetic. In the first subsection, the outline of the algorithm is described. The algorithm is based on Algorithm 2 and the transformation by which the tangent vector of solution curve and one axis has the same direction. In the second subsection, the region where the solution curve is expected to exist is described. In the last subsection, the proposed algorithm is shown.

**3.1. Outline of New Algorithm**

Let \( X \) be the space of \( x \in \mathbb{R}^{n+1} \) and let \( Y \) be another space of \( y \in \mathbb{R}^{n+1} \). Let us consider a transformation \( \phi: Y \to X \)
\[ x = \phi(y) = c + f'(c)^{tr}(y_1, \ldots, y_n)^{tr} + y_{n+1}v. \]
Remark that \( y = 0 \) corresponds to \( x = c \) and that the \( y_{n+1} \)-axis and \( v \) has the same direction.

Denote the point in \( Y \) by \((\cdot)_Y\) and denote the group \((y_1, \ldots, y_n)\) by \( y^{(n)} \). The point on the tangent vector \( v \) can be described as \((0, \ldots, 0, v)_Y\), with a parameter \( s \). In the proposed method, we obtain the crossing point between the solution curve and the hyper-plane which is through the point \((0, \ldots, 0, 1)_Y\) and which intersects orthogonally to the vector \((0, \ldots, 0, 1)_Y\). Thus, the original problem is translated into obtaining the point \( (y^{(n)}, s)_Y \) such that
\[ f(x) = f(c + f'(c)^{tr} y^{(n)} + y_{n+1} v) = 0. \]

Remark that the hyper-rectangle whose each boundary are parallel to the corresponding axis in \( Y \) is the one whose one boundary is parallel to the tangent vector \( v \) of the solution curve in the original space \( X \).

Assume that the test region is set as the Affine Form \((d \varepsilon^{(n)}, s)_Y\) with some \( d \in \mathbb{R} \), where \( \varepsilon^{(n)} \) denotes \((\varepsilon_1, \ldots, \varepsilon_n)\). The problem guaranteeing the existence of the solution \((y_1, \ldots, y_n, s)_Y\) of Eq.(4) in the region \((d \varepsilon^{(n)}, s)_Y\) is translated into the one that there exists \( \varepsilon^{(n)} \) in the \( n \)-dimensional unit ball \( B^{(n)} \) such that
\[ f(c + df'(c)^{tr} \varepsilon^{(n)} + sv) = 0. \]

Newton operator \( k(\varepsilon^{(n)}) \) for obtaining the solution \( \varepsilon^{(n)} \) of Eq. (5) is described as
\[ k(\varepsilon^{(n)}) = \varepsilon^{(n)} - Rf(c + df'(c)^{tr} \varepsilon^{(n)} + sv) \]
using some \( n \)-dimensional square matrix \( R \). Existence of the fixed point of \( k \) in the unit ball can be guaranteed by the condition
\[ \{ \varepsilon^{(n)} - Rf(c + df'(c)^{tr} \varepsilon^{(n)} + sv) \mid \varepsilon^{(n)} \in B^{(n)} \} \subset B^{(n)}. \]

Since we are subject to guarantee the existence of solution curve, that is, the solutions \( \varepsilon^{(n)} \) for all \( s \in [-3, 3] \), we must check the condition
\[ \{ \varepsilon^{(n)} - Rf(c + df'(c)^{tr} \varepsilon^{(n)} + sv + n_{n+1} v) \mid \varepsilon^{(n)} \in B^{(n)} \} \subset B^{(n)}. \]

The left-hand side of the condition (6) can be evaluated by Affine Arithmetic. In the next two subsections, how to determine \( d \) and \( R \) are described, respectively.

**3.2. How to obtain \( d \)**

In this subsection, the radius \( d \) of the region where the solution curve is expected to exist is determined. It is known that the optimal radius of the test region \( T \) in Krawczyk method is the double distance between the approximate solution \( c \) and the image of Newton operator for \( c \). This idea is used to obtain \( d \). Newton operator \( k_{y^{(n)}}(y^{(n)}) \) for Eq.(4) is described as
\[ k_{y^{(n)}}(y^{(n)}) = y^{(n)} - \frac{df(c + f'(c)^{tr} y^{(n)} + sv)}{dy^{(n)}} \cdot f(c + f'(c)^{tr} y^{(n)} + sv). \]
Since $0_Y$ is regarded as the approximate solution of Eq. (4), the distance between the approximate solution $0_Y$ and the image $k_y^{(n)}(0)$ of Newton operator $k_y(n)$ for $0_Y$ is described as
\[
\|0_Y - k_y^{(n)}(0_Y)\|_\infty = \left\| \frac{df((c + f'(c)y^{(n)})^{-1}y^{(n)})}{dy^{(n)}} \right\|_{y^{(n)}=0} f(c + sv) = \left\| (f'(c)f'(c)^{tr})^{-1} f(c + sv) \right\|_\infty.
\]

Since we now consider all $s \in [-\bar{s}, \bar{s}]$, that is $\bar{s}e_{n+1}$, Thus, the radius $d$ of the region where the solution curve is expected to exist is obtained as
\[
d = 2 \left\| (f'(c)f'(c)^{tr})^{-1} f(c + \bar{s}e_{n+1}) \right\|_\infty.
\]

3.3. How to obtain $R$

In this subsection, $R$ of (6) is determined and the left-hand side of (6) is transformed into another form for the effectiveness.

Using Affine Arithmetic, assume that we evaluate
\[
f(c + df'(c)^{tr}e^{(n)} + se_{n+1}) = a + Lc^{(n)} + \bar{L}e,
\]
where $a \in \mathbb{R}^n$, $L \in \mathbb{R}^{n \times n}$, $\bar{e}$ is the set of $(e_{n+1}, e_{n+2}, \ldots)$ by Affine Arithmetic for nonlinear operators and $\bar{L}$ is the factor of such $\bar{e}$.

By choose $R = L^{-1}$, the left-hand side (6) can be described as
\[
e^{(n)} - L^{-1}(a + Lc^{(n)} + \bar{L}e) = e^{(n)} - L^{-1}a - e^{(n)} - L^{-1}\bar{L}e = -L^{-1}a - L^{-1}\bar{L}e.
\]

Thus, the condition
\[-L^{-1}(a + \bar{L}e) \subset B^{(n)}
\]
can be used to prove the existence of the solution curve.

3.4. Algorithm

**Algorithm 3**

Step 1. Give $\bar{s} > 0$.

Step 2. Obtain $v \in \mathbb{R}^{n+1}$ such that
\[
\begin{align*}
&f'(c)v = 0, \\
&(v, v).
\end{align*}
\]

Step 3. Obtain $d \in \mathbb{R}$ such that
\[
d = 2 \left\| (f'(c)f'(c)^{tr})^{-1} f(c + \bar{s}e_{n+1}) \right\|_\infty.
\]

Step 4. Evaluate $f(c + df'(c)^{tr}e^{(n)} + \bar{s}e_{n+1})$ as the Affine Form
\[
a + Lc + \bar{L}e
\]
using Affine Arithmetic.

Step 5. Obtain $R = L^{-1}$.

Step 6. If
\[-R(a + \bar{L}e) \subset B \]
holds, there exists a solution curve in $c + df'(c)^{tr}e^{(n)} + \bar{s}e_{n+1}$. Otherwise, reduce $\bar{s}$ and go back to Step 3.

4. Conclusion

In this paper, we proposed a new algorithm to prove the existence of solution curves of nonlinear equations. The algorithm is based on Affine Arithmetic and the transformation by which the tangent vector of solution curve and a boundary of region where solution curve is expected to exist have the same direction. Thus, the algorithm is expected to be more effective than the conventional method in which such regions are described by intervals. Numerical examples will be shown in the presentation of this paper.

References


Evaluation of Route Expected Transmission Count of Routing Strategies in Multi-hop Wireless Networks

Kazuyuki Miyakita†, Keisuke Nakano‡, Masakazu Sengoku‡, Shoji Shinoda‡

†Dept. of Information Engineering, Niigata University, 2-8050, Ikarashi, Niigata 950-2181, Japan.
‡Dept. of Electrical, Electronic and Communication Engineering, Chuo University, Tokyo 112-8551, Japan.
Email: kmiya@net.ie.niigata-u.ac.jp, [nakano, sengoku]@ie.niigata-u.ac.jp, shinoda@elect.chuo-u.ac.jp

Abstract—In this paper, we consider to minimize the route expected transmission count (ETX) in multi-hop wireless networks. The route ETX is the predicted number of data transmissions required to send a packet over that route and is inversely proportional to the throughput. We theoretically compute the mean route ETXs of Shortest Path Routing and Longest Path Routing, and compare them with that of Optimum Routing, which always minimizes the route ETX. We show that the former two algorithms cannot minimize the mean route ETX if the density of nodes is large. To overcome this problem, we propose a routing algorithm which selects a path with the smaller number of hops than Longest Path Routing and shorter links than Shortest Path Routing. We show that this algorithm can approximately minimize the mean route ETX like Optimum Routing even if the density of nodes is large.

1. Introduction

In multi-hop wireless networks [1], [2], there are a lot of candidate multi-hop paths which can connect source and destination because of randomness of distribution of relay nodes and limitation of radio transmitting range. Usually, routing algorithms select a path in a shortest path fashion. Then, distance between relay nodes tends to be long and the loss ratio of a link is large. As a result, retransmissions frequently occur [3]. One way to reduce such retransmissions in each link is to choose shorter links than those selected in a shortest path fashion. However, choosing such shorter links increases the number of hops, and the increase of the number of hops means the increase of the total number of retransmissions over the path, which are the sum of retransmissions in all links included in the path. From such a viewpoint, we have a question as follows: which is the best path with the minimum number of retransmissions?

In this paper, we consider this problem by comparing some routing algorithms using a metric called the route expected transmission count (ETX). The route ETX is the predicted number of data transmissions required to send a packet over that route, including retransmissions [3]. Namely, the route ETX is the sum of ETXs of all links in the route. Throughput between source and destination is inversely proportional to the route ETX. In this paper, we assume that the ETX of a link is a function of length of the link to clarify the effect of randomness of distance between adjacent nodes on the route ETX. We also assume that this function is a monotonically increasing function. We analyze the route ETXs of some typical routing strategies in street multi-hop wireless networks.

We theoretically derive the mean route ETXs of Shortest Path Routing and Longest Path Routing. We compare the mean route ETXs of Shortest Path Routing and Longest Path Routing with that of Optimum Routing, which always minimizes the route ETX, and show that the former two algorithms cannot minimize the mean route ETX if the density of nodes is large. To overcome this problem, we propose a routing algorithm which selects a path with the smaller number of hops than Longest Path Routing and shorter links than Shortest Path Routing. We show that this algorithm can approximately minimize the mean route ETX like Optimum Routing even if the density of nodes is large.

2. Definitions and Assumptions

We consider a street multi-hop wireless network consisting of nodes randomly distributed along a street as shown in Fig. 1. Assume that the positions of nodes obey a Poisson process with intensity \( \lambda \), and all nodes are stationary. Let \( n \) be the number of nodes between source and destination. Let \( v_0 \) and \( v_{n+1} \) be source and destination, respectively. Denote by \( v_1, v_2, ..., v_n \) the nodes between \( v_0 \) and \( v_{n+1} \) as shown in Fig. 1, where \( n = 7 \). Let \( d \) be the maximum radio transmitting range of a node. Namely, two nodes are directly linked if distance between them is not longer than \( d \). Otherwise, they are not linked.

![Figure 1: Model.](image-url)

Let \( u(z) \) be the function of ETX of a link of length \( z \). The route ETX is the sum of ETXs of all links in the route. In
Fig. 1. $v_0$ is connected to $v_8$ via nodes $v_2$, $v_5$ and $v_6$, and $z_1$, $z_2$, $z_3$ and $z_4$ are lengths of links in Fig. 1. Then, the route ETX is computed as $u(z_1) + u(z_2) + u(z_3) + u(z_4)$.

3. Basic Routing Algorithms

In this section, we consider the following three basic routing algorithms. Longest Path Routing (LPR) selects a multi-hop path which includes all of the $n$ relay nodes as shown in Fig. 2(a).

Shortest Path Routing (SPR) selects a multi-hop path which includes the minimum number of relay nodes as follows: Consider Fig. 2(b). Let $z_i$ be the position of $v_i$. Let $l_i$ be the number of nodes in the interval $(z_i, z_i + d)$. SPR connects $v_0$ and $v_{l_i}$. If $v_i$ is selected as a relay node, $v_i$ is connected to $v_{i+1}$. In the same manner, a multi-hop path to $v_{n+1}$ is constructed.

Optimum Routing (OR) computes the ETXs of all paths and selects a path which minimizes the route ETX. OR is an ideal one, and it is difficult to use OR in actual situations; however, we evaluate the route ETX of OR to know the minimum value of the route ETX. LPR and SPR are extreme routing algorithms. LPR reduces retransmissions between adjacent nodes; however, it may increase the total number of retransmissions due to increasing the number of hops. SPR constitutes a reversal of LPR.

In the following, we theoretically analyze the mean route ETXs of LPR and SPR given that a source and destination have at least a multi-hop path.

3.1. Route ETX of Longest Path Routing

Define that $f_c(n)$ is the probability that the number of relay nodes is $n$. Define that $p_u(x, d)$ is the probability that source and destination are connected given that the number of relay nodes is $n$. Define that $F_h(x, d)$ is the probability that source and destination are connected. Then, from these probabilities and the density function, the mean route ETX for LPR, denoted by $U_L(x)$, can be computed as follows:

\[
U_L(x) = \begin{cases} 
\int_0^x u(z) e^{-ax} [d(x-z)] + e^{-ax} u(x) & (0 < x \leq d), \\
\sum_{n=0}^x f_c(n) (n+1) \int_0^x f_c(z) u(z) dz & (x > d)
\end{cases}
\]

where

\[
f_c(n) = \frac{(\lambda x)^n}{n!} e^{-\lambda x} \frac{p_u(x-d)}{F_h(x, d)},
\]

\[
f_c(z) = \frac{n}{x} (1 - z/x)^n - 1 \frac{p_u(x-z, d)}{p_u(x, d)},
\]

\[
p_u(x, d) = \sum_{j=0}^{[\frac{x}{d}]} (-1)^j \frac{(n+1)!}{j!(n+1-j)!} \left(1 - \frac{d}{x}\right)^n
\]

\[
F_h(x, d) = \sum_{i=0}^{[\frac{x}{d}]} (\lambda x)^i \left\{ e^{-\lambda x}[x-i] \right\}.
\]

where $[\frac{x}{d}]$ denotes the integer part of $\frac{x}{d}$.

3.2. Route ETX of Shortest Path Routing

Consider a case where $0 \leq x \leq d$. In this situation, $v_0$ is directly connected to $v_{n+1}$. Then, the mean route ETX for SPR, denoted by $U_S(x)$, can be computed as $U_S(x) = u(x)$.

Next, we consider a case where $d < x \leq 2d$. If $d < x \leq 2d$, at least one relay node is required to connect $v_0$ and $v_{n+1}$ because they cannot be directly connected. Suppose that there are just $m_1$ nodes in the interval $(0, d]$. Let $w_1$ be the position of $v_{m_1}$. If $x-d \leq w_1 \leq d$, $v_0$ is connected to $v_{n+1}$ via $v_{m_1}$, and the route ETX is $u(w_1) + u(x-w_1)$. The probability that two hops are sufficient to connect $v_0$ and $v_{n+1}$ given that $v_0$ and $v_{n+1}$ are connected, denoted by $\rho_{c2}$, can be computed as $\rho_{c2} = 1 - e^{-\lambda(2d-x)}$. The probability density function of $w_1$ given that two hops are sufficient to connect $v_0$ and $v_{n+1}$ and $d < x \leq 2d$ is $f_u(w_1) = \frac{e^{-\lambda x}}{\rho_{c2}}$. Then, the mean route ETX given that $d < x \leq 2d$ and two hops are sufficient to connect $v_0$ and $v_{n+1}$ can be computed as follows:

\[
U_{S22}(x) = \int_x^{d} f_u(w_1)(u(w_1) + u(x-w_1)) dw_1.
\]

If $0 \leq w_1 < x - d$, because this relay node is not connected to $v_{n+1}$, two relay nodes are required to connect $v_0$ and $v_{n+1}$.
Suppose again that there are just \( m_1 \) nodes in the interval \((0, d]\), \( w_1 \) is the position of \( v_{m_1} \) and \( 0 < w_1 < x - d \). Suppose that there are \( m_2 \) nodes in the interval \((w_1, w_1 + d]\). Then, we can connect \( v_0 \) and \( v_{m_1 + 1} \) via \( v_{m_1} \) and \( v_{m_1 + m_2} \). Let \( w_2 \) be the position of \( v_{m_1 + m_2} \). The probability that three hops are needed to connect \( v_0 \) and \( v_{m_1 + 1} \) given that \( v_0 \) and \( v_{m_1 + 1} \) are connected, denoted by \( p_{c3} \), can be computed as \( p_{c3} = 1 - e^{-d(x - d)} \). The probability density function of \( w_1 \) given that three hops are needed to connect \( v_0 \) and \( v_{m_1 + 1} \) and \( 0 < w_1 < d - x \) is \( f_{w_1}(w_1) = \frac{d}{w_1^{1/2} \pi^{1/2}} \). Also, the probability that \( w_2 \) is in the interval \((d, w_1 + d]\), denoted by \( p_{m}(w_1) \), can be computed as \( p_{m}(w_1) = 1 - e^{-d w_1} \). The probability density function of \( w_2 \) given that three hops are needed to connect \( v_0 \) and \( v_{m_1 + 1} \) and \( d < w_2 \leq w_1 + d \) is \( f_{w_2}(w_2) = \frac{d}{w_2^{1/2} \pi^{1/2}} \). Furthermore, the route ETX in this case is \( u(w_1) + u(w_2 - w_1) + u(x - w_2) \). Therefore, the mean route ETX given that \( d < x \leq 2d \) and three hops are needed to connect \( v_0 \) and \( v_{m_1 + 1} \) can be computed as follows:

\[
U_{x:3}(x) = \int_{0}^{x-d} f_{w_1}(w_1) p_{m}(w_1) \int_{d}^{w_1+d} f_{w_2}(w_2) \times \{u(w_1) + u(w_2 - w_1) + u(x - w_2)\} dw_2 dw_1.
\]  

(7)

From Eq. (5), (6), (7), the mean route ETX given that \( v_0 \) and \( v_{m_1 + 1} \) are connected for \( d < x \leq 2d \) can be computed as follows:

\[
U_3(x) = \frac{p_{c3}U_{x:2}(x) + p_{c3}U_{x:3}(x)}{F_h(x, d)} \quad (d < x \leq 2d).
\]  

(8)

For \( x \geq 3d \), we can compute \( U_3(x) \) in the same manner.

### 3.3. Results and Discussion

In this section, we numerically compute the mean route ETXs of LPR and SPR. In this computation, we use a model of \( u(z) \) as explained below. Assume that \( u(z) \) is a function of \( p(z) \), which is the probability that a packet is successfully transmitted over a wireless link of length \( z \) as \( u(z) = \frac{1}{p(z)} \) [3]. We use the following model as \( p(z) \), which is derived in [5] with assumptions that non-coherent FSK modulation and NRZ encoding are used.

\[
p(z) = \int_{-\infty}^{\infty} g(X_{dB}) \left(1 - \frac{1}{2} e^{-\frac{z}{2\sigma^2}}\right)^{8f} dX_{dB}.
\]  

(9)

In this equation, \( X_{dB} \) is a zero-mean Gaussian RV (in dB) with standard deviation \( \sigma \), \( g(X_{dB}) \) is the probability density function of \( X_{dB} \), \( f \) is the packet size (in bytes). Fig. 3 shows examples of \( p(z) \) and \( u(z) \) with the following parameters and \( y(z) \) assumed in [5]: \( \sigma = 4 \), and \( f = 50 \) bytes. We use these parameters in this paper. As can be seen in Fig. 3, \( u(z) \) rapidly increases around \( z = 23 \). Hence, we set \( d \) to \( 23 \).

Fig. 4 shows the numerical results of the mean route ETXs as a function of \( x \) for LPR and SPR for \( \lambda = 0.05 \) and \( 0.5 \). For \( \lambda = 0.05 \), the mean number of nodes in a transmitting range is equal to 1.15. Then, we have about one candidate path. Hence, the routing strategies cannot choose different paths. When \( \lambda > 0.05 \), we can have more candidate paths. Therefore, the routing strategies can choose different paths according to their policies. As a result, difference in the route ETX between the routing strategies becomes more remarkable as \( \lambda \) increases. Simulation results are also plotted to confirm the validity of the above analyses. From this figure, we can confirm that numerical results agree well with the simulation results. This figure also shows the simulation results of OR.

![Figure 3: Packet reception rate of a link \( p(z) \) and ETX of a link \( u(z) \).](image)

![Figure 4: Mean route ETX.](image)

For \( \lambda = 0.05 \), the mean route ETX for LPR is less than
for SPR for almost all $x$. Hence, for small value of $\lambda$, LPR is better than SPR. Furthermore, the mean route ETX for LPR is close to OR. This result comes from the fact that the number of candidate paths is not so much for small $\lambda$.

On the other hand, for $\lambda = 0.5$, the mean route ETX for SPR is less than for LPR for almost all $x$. Hence, SPR is better than LPR if $\lambda$ is large as opposed to small value of $\lambda$. Furthermore, there is a big difference between OR and SPR. This is because the number of hops increases for LPR and length of each link increases for SPR as $\lambda$ increases. Therefore, we have to consider another kind of policy different from shortest path and longest path fashions to minimize the mean route ETX.

4. Adjustable Routing

We propose another routing strategy which selects a path including shorter links than a shortest path and less relay nodes than a longest path. We call this strategy Adjustable Routing (AR). Consider Fig. 2(c). Define that $d_i$ is a positive constant which is not larger than $d$. Let $z_i$ be the position of $v_i$. Let $\ell_i$ be the number of nodes in the interval $(z_i, z_i + d_i]$. AR connects $v_0$ and $v_{\ell_0}$. If $\ell_0 = 0$, $v_0$ is connected to $v_1$. If $v_i$ is selected as a relay node, $v_i$ is connected to $v_{i+\ell_i}$. If $\ell_i = 0$, $v_i$ is connected to $v_{i+1}$. In the same manner, a multi-hop path to $v_{\ell+1}$ is constructed. This routing aims to make length of a link shorter than $d_i$ and make the number of hops close to $\frac{d}{d_i}$. We will explain the way to decide $d_i$ to minimize the route ETX in the following section.

4.1. Determination of $d_i$

This subsection proposes a method to determine the value of $d_i$ for AR assuming that $u(z)$ is a convex monotonically increasing function. Consider two nodes $v_i$ and $v_j$. Let $y$ be distance between $v_i$ and $v_j$. Then, the ETX of the direct link between $v_i$ and $v_j$ is equal to $u(y)$. Suppose that there is a node $v_k$ between $v_i$ and $v_j$. Let $y_1$ and $y_2$ be distance between $v_i$ and $v_k$ and that between $v_k$ and $v_j$, respectively. Then, the ETX of the path consisting of $v_i$, $v_k$, and $v_j$ is equal to $u(y_1) + u(y_2)$. If $u(y) \leq u(y_1) + u(y_2)$, direct connection between $v_i$ and $v_j$ realizes smaller value of ETX than two hop connection via $v_k$. $u(y)$ is always smaller than $u(y_1) + u(y_2)$ if $u(y) \leq 2u\left(\frac{y}{2}\right)$ under the assumption that $u(z)$ is a convex monotonically increasing function. Let $d_i$ be the maximum value of $y$ which satisfies $u(y) \leq 2u\left(\frac{y}{2}\right)$. In AR, we try to connect two nodes directly if $u(y) \leq 2u\left(\frac{y}{2}\right)$ to prevent a long link with poor ETX. Namely, we connect two nodes directly if distance between them is less than $d_i$. Furthermore, to prevent increase of the number of hops, we try to make length of a link close to $d_i$.

4.2. Result and Discussion

In Fig. 4, the mean route ETX of AR is plotted with those of SPR, LPR and OR. The mean route ETX of AR is computed by computer simulation. As a function $u(z)$, we use the function shown in Fig. 3. Then, we obtain $d_i = 16.9$ by using the proposed method. To confirm that determination of $d_i$ is properly done, we show the mean route ETX of AR with other values of $d_i$. For example, we show the results with $d_i = 8.9, 12.9, 16.9, 20.9$ in the figure. As shown in Fig. 4, for both $\lambda = 0.05$ and $\lambda = 0.5$, the mean route ETX of AR with $d_i = 16.9$ is the smallest in those of AR with the above parameters. Also, the mean route ETX of AR with $d_i = 16.9$ is almost the same as that of OR even in the case of large $\lambda$. From these results, we can confirm that AR with $d_i$ determined by the proposed method works well.

5. Conclusions

We have analyzed the mean route ETXs of four routing algorithms. We have theoretically computed the mean route ETXs of Longest Path Routing and Shortest Path Routing. We compared these numerical results of the theoretical computations with the mean route ETX of Optimum Routing, which is computed by computer simulation. From the comparison, we have showed that, if the density of nodes is small, the mean route ETX of Longest Path Routing is smaller than that of Shortest Path Routing and, if the density is large, the mean route ETX of Shortest Path Routing is smaller than that of Longest Path Routing. We have also showed that difference between the mean route ETX of Optimum Routing and that of Shortest Path Routing is large if the density of nodes is large. To minimize this difference, we have proposed a simple routing called Adjustable Routing. Adjustable Routing makes the number of hops smaller than Longest Path Routing and makes length of each link in a multi-hop path smaller than Shortest Path Routing. From the computer simulation results of the mean route ETX of Adjustable Routing, we have showed that Adjustable Routing realizes almost the same mean route ETX as Optimum Routing. As an application, the concept of AR can be used to select a route in mesh networks where distance between nodes can be estimated.

References

Channel features observed to model the long term behavior of wireless links

Eugenio Costamagna, Lorenzo Favalli, and Pietro Savazzi

Department of Electronics, University of Pavia
via Ferrata 1, 27100 Pavia, Italy
Email: name.surname@unipv.it

Abstract— Chaotic dynamical equations have been introduced as an effective tool to model channel behaviors. Various channel statistical features were used to derive the model parameters and then to generate realistic realizations of the error process. In this paper, we attempt to discuss the influence of the so called variation coefficients on the medium and long term behavior of the error sequences, making comparison with results obtained from the optimization of the logscale diagram. We show that both criteria are suitable to obtain good results, considering the performances derived from the error sequences at the higher level of ARQ receiving procedures and at the output of a soft decoding process.

1. Introduction

Chaotic dynamical equations have been introduced to model time series derived from transmission channels, considering both the error process for fading channels [1][2], and the sampled attenuation process for satellite channels in the Ka band: see for instance [3] and references quoted therein. The modeling procedure optimizes a weighted sum of the so called strange attractors as described in [1], which reads:

\[ S = F[\sum k_i u_i]. \]  \hspace{1cm} (1)

The \( u_i \) variables are the projections of the attractor trajectory point vectors on the \( x \) coordinate, or a similar geometric characteristic. The \( k_i \) are suitable weights, and index \( i \) is ranging from 1 to the number of used attractors, normally 3 to 7. \( F(\cdot) \) is a probability shaping function, matching the distribution of the sample amplitudes derived from (1) to the target time series and \( f(\cdot) \) is a polynomial or exponential function providing the suitable output range. Function \( F \) is obtained by means of a suitable inversion procedure for the cumulative probability function [1].

Both the weights and the trajectory sampling distances are optimized to match the characteristics of the target sequence by means of tabu search metaheuristics, or by the Nelder and Mead simplex procedure, which is used in the present work. Suitable cost functions have been introduced [1-3] to obtain behaviors mimicking the target sequence, considering block error probabilities \( P(m,n) \) (being \( m \) the numbers of errors in bit blocks of length \( n \)), moving autocorrelation and autocovariance functions and histograms, mean differences between values sampled at various distances, and the so called variation coefficients \( K(r) \). The last, already discussed in [4], relate to the variances of the so called multigaps, i.e., the sums of \( r \) consecutive gap lengths, accounting for the renewal or non renewal character of the error process. They are perhaps the only truly long term features, to which attention is focused. Good bit error rates and gap length probabilities are built in the process via to above shaping: on the whole, the approach is somewhat complementary with respect to other recent works based on error statistics, see for instance [5].

In this paper, we try to evaluate the relevance of the variation coefficients from the point of view of the performance of an ARQ (Automatic Repeat reQuest) transmission protocol or at the output of a decoding process. This way, the statistical feature controlling the generation of error burst is related to its consequences at an higher level of observation, and the relevance of using it in the optimization procedure is clearly shown.

Hidden Markov models (HMM), optimized by means of the Baum-Welch procedure, or Gilbert’s models are introduced for the sake of comparison. Normally, these models supply flat \( K(r) \) curves, different from those exhibited by burst channels. Instead, the chaotic models described in [1-3] can be easily matched to a target sequence and, equally well, optimized to produce sequences exhibiting all the same statistics but different \( K(r) \)s. So, we exploit this capability to enlighten the effect of changing only the behavior of the \( K(r) \)s, mimicking that derived from a HMM in place of the target one.

We consider two target sequences of error gaps, derived by means of physical layer simulations: the first is obtained from a Rayleigh 3 paths channel, with a maximum Doppler frequency of 120 Hz and a delay spread equal to half the symbol rate, typical for urban or indoor wireless channels; the bit error rate (BER) is about 7.2x10^{-2}. The second is a post detection bit error stream derived from a soft decision turbo decoder, with a BER of about 4.1x10^{-3}. As physical layer simulators are computationally heavier than chaotic and HMM models, effects, they are usually used for tuning faster simulators like the ones discussed in this paper.

From the target and the modeled gap series we derive sequences of accepted or rejected blocks of bits, according to a suitable ARQ protocol. Then, we examine the correlation properties of the sequences of gaps between rejected blocks (i.e., numbers of consecutive accepted blocks). Later, we optimize logscale diagrams [6] instead of variation coefficients: the agreement between the long term behaviors of the sequences.
obtained with the two approaches is noticeable, and we derive some important observations.

2. Results for the Rayleigh channel

The sequence length is about 145,000 gaps and a HMM sequence of the same length was provided.

Good results have been obtained with seven attractors for all the chaotic models described in the following, looking for both small and large P(m,n) values, moving autocorrelation and autocovariance functions, moving histograms and average value differences sampled at 20, 100 and 1000 gap distances. For the P(m,n) values and the K(r) values we follow [1] with block lengths ranging from 10 to 80, and we introduced in the cost function a further error term to optimize the derivatives dK(r)/dr together with the K(r) values. We optimized three chaotic models. The first is matched to all the target features. The others mimic all the same features but the variation coefficients, in place of the behavior of the target sequence, high and very high K(r) values are imposed. As K(1) indicates the extent of error clustering that the process exhibits compared to total randomness, and K(2) - K(1) is proportional to the correlation between successive gaps (see [4], paragraph V-D), we were looking for very accurate values at least for the very first values of the K(r) curve. So, the values obtained for the first curve and for the low K(r) curve agree with the target for r < 5 within a few percent. Autocorrelation and moving autocovariance functions and histograms are evaluated as in [1-3].

As the K(r) curve exhibited by the target is relatively flat, the Baum-Welch procedure is effective in this case in producing a HMM, with 5 good states and 3 bad states, leading to good results not only for the other statistics, as usual, but for the variation coefficients too: only the very first values are somewhat high. No better performances were obtained with larger numbers of states are allowed.

We plot in Fig. 1 the results obtained from long runs of the models (about 400,000 gaps, i.e., about 5 Mbit).

Due to the flexibility of the optimization procedure of the chaotic models, behaviors very similar to the target have been saved for all the other features. So, it is expected that the differences perceived at the output level have to be referred only to the different variation coefficient behavior of Fig. 1.

Now, let we examine the effects on the ARQ transmission. For this, considering realistic propagation times in transmission, it will be sufficient to consider the retransmission rate and the occurrence of rejected blocks of bit, and we make reference to a block length of 80 bits. For this length, the P(m,80) curves obtained from all the five sequences used as inputs were very similar: assuming a threshold of m=2 errors per block and computing the probabilities of having more than 2 errors any block of 80, the agreement was within a few percent.

In Fig. 2 we observe the behavior of the autocorrelation functions of the block rejection processes (normalized to 1 at lag 0): the target and the matched sequences, as well as the HMM, are not exhibiting long term effects, whereas these are well evident in the curves obtained from the sequences exhibiting high and very high variation coefficient values. Notice also the secondary peaks for large lags for the high and very high K(r) curves. An autocorrelation function a little lower than the dotted line in Fig. 2 was obtained from a sequence exhibiting the same characteristics of the high K(r) values, one, except somewhat lower K(r) at the right hand of Fig. 1, starting from about r = 600. This is clearly showing the influence of the behavior of the variation coefficients not only on the short term gap correlation, but on the long term too. The 20 blocks lag, in which differences are perceived in Fig. 2, spans over 1600 bits, i.e., a distance for which the autocorrelation functions for the gap and for the bit processes were not revealing notable differences. In Fig. 3 we report the probabilities of obtaining repeated block rejection. Block retransmission is performed following a go-back-N ARQ protocols, with lengths N used as abscessas.
Fig. 3. Percent probabilities of obtaining repeated errors exceeding the same threshold used in Fig. 6 (2 error bits in any 80 bit block) during go-back-N ARQ transmission. N is the abscissa block lag. Symbols as in Fig. 2 (dotted and dashed-dotted lines are near coincident).

Only the very high variation coefficient sequence shows a behavior very different from the others, being the HMM curve the best matched to the target, and differences among the curves are small. Experience shows that the behaviors of the curves obtained from sequences with high variation coefficients is somewhat erratic: sometimes they are similar to the target curve, sometime similar to curves obtained from very high $K(r)$s.

Similar results have been obtained for a block length of 47 bits and only 1 error accepted in any block, with sequences derived from long runs of the generators too (up to 400,000 gaps).

3. Results for a postdetection error stream derived from a soft decoding process

A confirmation of the above results was obtained observing the error process at the output of a turbo decoder supplied with a soft bit stream generated by a physical layer simulator.

The reference error sequence is 20,454 gap long (about 5 Mbit), with BER = 4.12x10^{-3}, and a fast rising variation coefficient curve is observed. We repeat the calculation pattern of the previous paragraph to compare its statistics to those of a chaos sequence with matched variation coefficients and to those of a gap sequence obtained from a Gilbert model exhibiting a low and flat $K(r)$ curve. Again, except for the $K(r)$, all the statistics obtained from the models can be considered in good agreement with the target, even if a less satisfactory result was obtained for the moving histograms of the Gilbert sequence. Again, close agreement with the target was observed for the first $K(r)$ values obtained from chaos sequences.

We show the variation coefficients in Fig. 4. The $P(m,n)$s are similar to the target, even if for both models not so good as were for the Rayleigh sequences. However, the agreement remains satisfactory in the used region, where error probabilities are not very low.

For instance, the $P(1,80)$ figures are 0.142 for the target, 0.141 for the chaotic model, and 0.098 for the Gilbert’s model. A sample of the time behavior of the sequences is reported in Fig. 5, where the 1,000 gap interval shows impressively the burst character of the target and of the matched $K(r)$ sequences, and the more uniform character of the Gilbert’s. Neglect for the moment the third curve, to be discussed later.

Using the outputs of the decoder, for simulating again an ARQ process, we found, see figures 6 and 7, behaviors similar to those observed in figures 2 and 3. Considering again the sequence of inter rejection intervals, high values of the autocorrelation function are obtained from the target and from the matched-$K(r)$ chaotic sequence, but not from the Gilbert model, and only the matched chaotic sequence lead to ARQ performances similar to the target. The block length is again 80 bit, and the error threshold is now no more than 1 error accepted for block. These results are well confirming the behavior of the $K(r)$ curves observed at the post-detection level in [2].
Fig. 6. The autocorrelation functions of the error block rejection process for the soft bit block sequences when a maximum of 1 error is accepted in any 80 bit block. Solid line: the target; dashed line: the matched $K(r)$s sequences, showing behaviors similar to the target. Circles: the Gilbert sequence, showing low correlation.

Fig. 7. Percent probabilities of obtaining repeated errors exceeding the same threshold used in Fig. 6 (1 error bit in any 80 bit block) during go-back-N ARQ transmission. N is the abscissa block lag. Symbols as in Fig. 6.

Fig. 8. Rayleigh channel: the variation coefficients of the sequence with optimized logscale diagram (dashed line) compared to the target (solid line) and to the already matched ones (dashed-dotted line, as in Fig. 1).

Conclusions

Sequences derived from various chaotic and Markov models to simulate the behavior of target sequences but exhibiting different variation coefficients lead to different behaviors at the output of higher level processes. The influence of the variation coefficients has been clearly depicted: when they are well matched to the target, the output process is similar to that obtained from the target sequence, and when they are not matched, the process is different, independently from using chaotic or HMM models. The logscale diagram is suitable to substitute the variation coefficient information, leading to similar results. In fact, by means of optimizing it, good variation coefficient behaviors have been obtained.

4. Optimization of the logscale diagram and comparison of results

The logscale diagram is described in [6], with its capabilities in describing long term features of sample series, in particular self-similarity and long range dependence. We computed or optimized it only to evaluate its capability to lead to results similar to those obtained from the variation coefficients. The optimization is performed by introducing in the cost functions a term carrying the differences between target and tentative logscale diagrams, squared and summed up. The results for the Rayleigh channel are shown in Fig. 8, and confirmed by the third curve in Fig. 5.

In general, at a very first sight, it seems that optimizing logscale diagrams and optimizing variation coefficients is not very different: a successful optimization of the logscale diagram seems to secure a good variation coefficient behavior, and a successful optimization of the variation coefficients leads to logscale diagrams looking promising. Optimization of both features at the same time was possible and perhaps convenient.

References

Self-adaptive timing-synchronization ability in ad hoc networks based on IEEE 802.11 MAC protocol

Hisa-Aki Tanaka†, Osamu Masugata†, Daisuke Ohta†, Akio Hasegawa‡, and Peter Davis‡

†Department of Electronic Engineering, The University of Electro-Communications (UEC), 1-5-1 Choufugaoka, Choufu-shi, Tokyo 182-8585 Japan
‡ATR Adaptive Communications Research Laboratories 2-2-2 Hikaridai, Keihanna Science City, Kyoto 619-0288 Japan
Email: htan@ee.uec.ac.jp

Abstract—In multi-hop ad hoc networks that use IEEE 802.11, long transient resynchronization states are often observed numerically when multiple IBSS merge. We propose a simple modification of the conventional timing synchronization method to reduce such synchronization bottlenecks. When the proposed modification is applied, a self-adaptive synchronization ability is observed in simulations, which makes resynchronization times much shorter and reduces energy consumption.

1. Introduction

The ad hoc mode of IEEE 802.11 standard (hereinafter abbreviated as 802.11) supports the formation of independent basic service sets (IBSSs) in the absence of an access point. Based on the 802.11 media access control (MAC) protocol [1], large ad hoc networks are now being developed. In addition, as the nodes (mobile PCs or PDAs) become smaller, more efficient, power-saving control is required, and the transmission ranges of nodes should be limited if possible. In such networks, the timing synchronization function (TSF) is one of the essential integral components, since timing synchronization is required for frequency hopping spread spectrum and synchronous power-saving control. In this 802.11 TSF, each node transmits a beacon that carries timing information only when the node is elected through the contention of neighbouring nodes.

This contention is based on a random process and supposed to provide a fair election of the beacon transmission node. Most conventional studies on this timing synchronization problem have assumed closed, isolated networks within a single IBSS [2,3]. In contrast to such static cases, timing synchronization in dynamic environments is currently being considered (e.g., for the cases of merging two IBSSs [4] and of resynchronization in large, multi-hop networks [5]). In the present study, we consider multi-hop networks in such dynamic situations, and propose a simple modification of the 802.11 TSF. Interestingly, this simple modification greatly reduces the transient resynchronization time, especially for larger networks with hundreds of nodes that maintain a relatively small total number of awake nodes. The reason for the efficiency of the present method is also investigated by systematic simulation and analysis, revealing a kind of self-organized adaptive process that enables fast resynchronization.

2. Synchronization bottlenecks

We consider the simple, but generic examples shown in Fig. 1. Systematic simulations of the synchronization process were carried out by changing the number of nodes as well as the transmission range (R) of nodes.

As shown in Fig. 1(a), 143 synchronized nodes (denoted by ⋄) belonging to the same IBSS are randomly placed in a 220 m×220 m square, and a node (denoted by ◦) belonging to another IBSS with a later TSF timing is then moved to a certain position. Even though this node is fixed, its transmission range can temporally become wider. This has the same effect as the above case of node movement. Both of the above situations initiate a resynchronization
process until all of the nodes (144 nodes in this example) eventually become resynchronized to the later timing (\(\omega\)). Since this resynchronization time depends on the transmission range of the nodes as well as on the position of the introduced node, we systematically change the transmission range (\(R\)) from 24 m to 360 m (by 4-meter intervals) and 10,000 trials are carried out for each \(R\) with a randomly positioned joining node. To consider the worst-case scenario of the synchronization process, the power-saving timing of the introduced node is initially set 180 degrees out of phase to that of original nodes in these simulations. Namely, the introduced node is awake when most of (but not all of) the other nodes are in the power-saving state. Note that only awake nodes can receive beacons, and a node remains in its awake state for one beacon period (~100 ms: awake/power-saving cycle) directly after the node has transmitted a beacon by beacon contention (BC). This is the reason why the resynchronization process becomes complicated and often requires a long time.

As a limiting case of Fig. 1(a), we also consider the case of merging two IBSSs (cells), as shown schematically in Fig. 1(c). This situation arises when two small rooms (in which any two nodes can directly communicate) are joined by a short corridor.

In order to clarify the essential mechanism of the beacon propagation, the present simulations assume that node movement can be neglected in the resynchronization process. In addition, the following simplifying assumptions are made: (1) the transmission range is time-constant and uniform over the nodes, (2) the transmission delay of the beacons carrying the TSF timing is negligible compared to a single beacon slot (50\(\mu\)s in this case), and (3) small mismatches in clock frequencies are negligible compared to the time scale of the awake/power-saving cycle (beacon period), as a first approximation. Later, we consider factors (2) and (3) to be non-negligible and analyze their effects.

The obtained simulation results are summarized in Fig. 2, in which the averaged resynchronization times are plotted against the transmission ranges (data plotted with \(\times\) in Fig. 2(a)). This data set is obtained using a regular array of 144 nodes. For other node configurations, including the totally random case and the randomly disturbed array case, we have observed a pattern similar to that shown in Fig. 2(a). For the case in which the range of \(R\) is larger than 140 m, as shown in Fig. 2(a), a synchronization bottleneck emerges. This long resynchronization time can be explained by a spatio-temporal pattern of beacon propagations and frequent beacon collisions, as shown in Fig. 1(b) (for details, see [5]). In other words, this bottleneck comes from the combination of the multi-hop network structure and the beacon contention (BC) mechanism in the 802.11 TSF. We analyzed this BC process in detail when the bottleneck emerges and found that the fairness in beacon transmission is lost because nodes near the boundary of the network can transmit beacons more frequently than other nodes. This is because the 802.11 TSF adopts a local BC,

which is no longer fair when the local node density is not uniform over the network.

3. A simple method to reduce synchronization bottleneck

Since loss of fair beacon transmission causes synchronization bottleneck, an intuitive method of bottleneck reduction is to regulate the beacon transmissions of nodes near the boundary. In the present simulations, nodes with larger BT values near the boundary are selected, whereas other nodes are selected with smaller BT values. This is because BC becomes less competitive near the boundary. Therefore, we should examine what happens when each node regulates the beacon transmission for the case in which the BT is larger than a certain threshold. In order to examine this idea concretely and ensure that it is logically consistent with the 802.11 TSF, the following modifications are introduced: (i) For a node selected by BC, the beacon transmission is cancelled if its BT exceeds a given threshold. (ii) Such a node maintains its awake state over one beacon period directly after the cancellation of the beacon transmission.

In Fig. 2, we compare three numerical data sets (plotted by +, * and □ symbols) obtained using the proposed method to that obtained using the original 802.11 TSF (\(\times\)). These three data sets are obtained by averaging over 10,000 trials when the BT threshold is set to one, two, and three beacon slots, respectively. It is noted that even in the worst instances of these three data the synchronization time is shorter than that of the 802.11 case. In Fig. 2(a) (and
(Fig. 2(d)), the data plotted by the + symbols show a performance that is several times better than the 802.11 TSF (plotted by the × symbols), except for certain short transmission ranges ($R < 60$ m). In these simulations, the number of awake nodes is also analyzed systematically both for the resynchronization process (Fig. 2(b)) and for the static states after the resynchronization has been completed (Fig. 2(c)), where $R$ is set to 240 m. In Fig. 2(b), all three data sets (+, *, and □) have a larger number of awake nodes (averaging approximately five awake nodes) than that for the 802.11 (×). However, the total number of awake nodes required until the resynchronization is complete becomes smaller in these three cases, because the resynchronization time is much shorter than in the 802.11 case. In Fig. 2(c), the number of awake nodes is approximately the same (∼ one node on average) for $R$ larger than 180 m. In addition, in the proposed method, the total number of beacon transmissions is always less than that of the 802.11 TSF, which is a direct consequence of the assumptions (i) and (ii). Thus, the total energy consumption in the proposed method is expected to be smaller than that of the 802.11 TSF.

4. Bottleneck reduction mechanism

In this section, we analyze the reduction of the resynchronization bottleneck in the proposed method, and the observed efficiency is explained. In Fig. 3, the number of remaining, non-resynchronized nodes (● in Fig. 1) is plotted with respect to time for the four cases (+, *, □, and ×) corresponding to the data shown in Fig. 2. Initially, there are 143 nodes, and the number of nodes decreases monotonically as the synchronization proceeds. In the 802.11 TSF a long bottleneck always emerges. In contrast, in the proposed method, such a bottleneck is reduced (+, *, and □ symbols in Fig. 3), where a rapid resynchronization emerges in several dozen nodes simultaneously, leading to total resynchronization. This rapid resynchronization, which is a characteristic feature of the proposed method, is a byproduct of modifications (i) and (ii) above. A brief explanation of this mechanism is as follows. (a) As the resynchronization proceeds, the density of the non-resynchronized nodes is gradually lowered. (b) For such a low node density, BC becomes less competitive and modifications (i) and (ii) often cause some of the nodes to remain awake simultaneously. (c) In this case, if some of nodes within another IBSS (′) transmit beacons, then they are received by these non-resynchronized nodes, whereupon resynchronization is completed in a single beacon cycle.

Thus, the bottleneck reduction in the proposed method is realized by a kind of self-adaptive, rapid synchronization process and its energy consumption is also found to be efficient.

5. Conclusions

Based on the understanding of the synchronization mechanism in multi-hop ad hoc networks, a simple bottleneck reduction method is proposed. This method utilizes a self-adaptive, rapid synchronization process and its energy consumption is also found to be efficient.

Acknowledgments

This work was partially supported by NICT and the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 16760292 (2004), and stimulated by the activities of the AdHoc Wireless Network Consortium.

References


Abstract—In this paper a new completely distribute architecture for collaborative environment is proposed. We consider the problem of distributing live streaming media content without any dedicated infrastructure and the aim of the platform is to provide a distributed environment in which every peer will be able to share all the media required in collaborative sessions i.e., synchronous e-learning sessions, video-conferences… Our P2P approach allows to scale the required bandwidth with demand (the amount of interested client) and allow the self optimization of the overlay tree structure which makes use of the Internet as a low level infrastructure to provide multimedia service to end host.

We point out that self-optimizing structure and robustness to peer transience are pivotal goals in collaborative environment in which participants can interact with each other for a long time and some of them can leave the session unexpectedly. In the following we present a distributed tree (therefore Peer tree Peer) management protocol that provide QOS optimization, peer failure recovery and, finally, we evaluate it by simulation.

1. Introduction

Due to the lack of wide spread support for IP Multicast, the simultaneously distribution of the same multimedia content to a potentially large group of client requires a set of resources that are not commonly available to everyone.

The target scenario of our real time collaborative environment are made by participants that would establish a multimedia session, for example, from a foreign network i.e., a network of an organization which is not the one in which the participant works or studies, and hence, the participant often has to use a slow or not very powerful hardware such as an old laptop. In addition, collaborative session could be infrequent and irregular and could not justify the cost of a server that should be sized to serve the maximum amount of client.

Today overlay networks, i.e. networks that uses Internet as a low level infrastructure, can be an alternative to IP multicast. A P2P approach avoids single participant resource limitation by scaling the request resources with demands: Every multimedia content, produced by a single Peer, could be delivered to all the other ones through an overlay tree structure like in many other overlay multicast structures [1], [2]. The tree structure can bound the resource requested to single peer but introduces a great limitation: a single peer failure compromises the service delivery to a potentially large peer subset. Using network path redundancy is a possibility that becomes very attractive with the introduction of Multiple Description Coding, [3], a method of encoding an audio and/or video signal into separate streams, or descriptions, such that any subset of these descriptions can be received and decoded. The distortion with respect to the original signal is commensurate with the number of descriptions received.

The rest of this paper is organized as follows. In section 2 we briefly explain the architecture design. The distributed management algorithm of every single tree of the structure is presented in section 3. In section 4 we start to evaluate the platform performances, finally, in section 5, we summarize the obtained results.

2. Architecture Overview

The entities of P3P architecture are Peer and User Agent (UA). The UA is the human front-end that transmits and receives all the multimedia streams; the peer, which is a member of the multicasting tree, act as a proxy from the UA point of view, and serves a small amount of UAs delivering to them the streams received in the multicasting tree, figure 1. Peer and User Agent are not necessary separated and could collapse in a single object, but dividing them permits us to abstract the overlay multicast structure from the specific front-end application and to take advantage of local IP multicast infrastructure.

The mean of the local scenario, the one in which all the
UAs are connected to the same peer, is to quickly connect a small amount of participants without increasing the global structure load: all the UA connected to the same peer communicate directly to each other and, if a UA is a stream source, it will send a copy of the captured streams to the peer that will propagate them into the multicasting tree.

No dedicated infrastructure means no rendez-vous server. The target application involves a relatively small amount of client and we can suppose that participants that collaborate in a project well know each other, hence, following the bit-torrent example a small .p3p file is published over a well known personal web site or could be sent by mail. The obtained .p3p file contains essentially the address of a remote peer of the multicasting tree, that have to be contacted from the local peer. After that the local peer will automatically integrate itself in the overlay multicast structure.

Network path redundancy could be obtained using multiple description coding techniques and delivering different description over different trees. Therefore a peer failure does not compromise all the description required from a potentially large sub-set of peer.

3. Tree Management

We discuss the problem of constructing and maintaining a single description distribution tree. In P3P the tree structure grow and try to optimize itself, at first, through a greedy algorithm and, hence, using a distributed local search optimization.

We assume that all the nodes participates and contribute bandwidth resources until the end of the session or until a failure.

Each peer compares the other one characteristics by using two evaluation metrics that measure the goodness of a peer as a parent node or as a child node. Parent metric is based on the evaluation of peer level i.e., the amount of intermediate nodes between the tree root and the parent node candidates. A peer is better than another one if his level is lower than the others. Peers with the same level are sorted by the measured delay between remote peer and the one that evaluates the metric.

Child metric is based on the outgoing bandwidth published by the remote peers. As in the Parent metric, peers with the same outgoing available bandwidth are sorted by the measured delay.

The algorithm try to reduce the height of tree structure i.e., it minimizes the amount of intermediate nodes between the root and the leaves. Shortness would minimize delay, section 4.1, and the probability of disruption due to peer failure. Each node selects his own children evaluating the Child metric: an incoming request will be accepted only if the metric of the peer that starts the Join procedure is greater than the minimum metric calculate over all the current children. Each node that searches a parent, sorts the peers that could become his own parent node by evaluating Parent metric over all the parent node candidates. With this strategies nodes with high outgoing bandwidth, hence, potentially a lot of children, gain a low level and the tree automatically becomes as bushy as possible.

When a parent receives and accepts a Join Request and the maximum amount of children (fixed by the maximum outgoing bandwidth) is reached, the worst child has to be removed. In order to avoid services interruption the child has not to be rejected until it finds a new parent node that starts to forward the required services. In section 4.2 we show the effectiveness of this soft handover strategy.

Each node constructs a local image of the overlay network based on the message received from other peer and, in this image, each remote peer is been identified as children, parent or other particular condition due to the previous communication. All the peer decision are based on the stored data that include all the remote peer published characteristics. Obviously, mistakes in the local image leads to unforeseeable effects. Remote peer are classified into the following states: probation a peer which has replied to a probing (discovery) message, parent candidate a peer which has not replied to a join request yet, parent a peer which has accepted the join request, future parent a peer which has frozen the join request, old parent a peer that was parent node but it is not the current parent, bad parent a parent node that has sent a remove injunction, child a child node, frozen child a future child node whose request could not to be accepted until a rejected child node finds a new parent, rejected child a child node that have to be removed.

3.1. Probing or Discovery phase

When a peer Q can not join remote peer it starts the discovery phase, figure 2. Discovery phase refresh the local structure image. Q sends a discovery message to his own parent that forwards the message to other peers following the previously established tree structure and, then, waits a timeout expiration.

Discovery phase is also performed when Q reads the .p3p file that contains the address of one of the node of the structure: Q knows only the address of one node, sends it a discovery message and all the peer which are interested to
become Q node parent send back to Q a reply. The reliability of the communication is not guaranteed but, if there are no packet loss, all the peers but the ones in the Q root sub-tree receive the discovery message. All peers that receive a discovery message forward it but reply to Q only if the characteristics of Q are better than the characteristics of their own worst child. The Discovery reply is sent directly to Q. When Q receives it he learns that the sender is not in his sub-tree and that, if a join request is sent, it will be accepted by the reply sender. At the timeout expiration Q evaluates the best parent candidate from the peers whose state is probed by using Parent metric. Discovery reply is also used to measure the path delay between peers. Discovery phase is also periodically performed when the peers try to optimize their own position in the tree structure.

3.2. Join phase

When Q chose the best parent candidate, R, it starts the join procedure and sends a join request to R, figure 3. If a timeout expires without any R reply the request is repeated. When R receives the request it evaluates the Q properties and sends an accept, freeze or reject reply. If Q child metric is smaller than the R child node smallest metric the reply is a reject one else, if R can serve immediately Q, it is an accept one else it is a freeze reply. When Q receives the R reply, it changes the R state, respectively, to parent and sends a positive or negative join confirm to R, future parent and waits for a short timeout expiration and then repeats the join to the same parent candidate or old parent and selects another parent candidate and repeat the Join phase. If there is no parent candidate node Q repeats the discovery phase. If the R reply was an accept reply R waits until a short timeout expiration hence repeats the positive reply; when it receives the join confirm, if positive, it allocates the needed resources and starts to forward the requested stream to Q which becomes a child node. Resource allocation have to be atomic operation hence only a new join has to be completed before a new one can start. After the join is completed R evaluates the best frozen child and eventually performs a new Join phase sending an accept join reply message. If the reply was a freeze reply R immediately sends a remove injunction message to his worst child, S. When S finds another parent node, R immediately notify to its best frozen child, that is not necessarily Q, an accept reply.

3.3. Reject phase

If R receives a request from Q and the Q child metric is greater than the smaller R child node metrics, hence the worst R child, S, has to be replaced by Q. R modifies the Q state to frozen child and the S one to rejected child. Soft handover policy imposes that S has to find a new parent node before the R reallocation resources. Hence R repeats periodically a remove message until S reply to him sending a new parent message.

When S receives a remove message from his parent node it modifies the R state to bad parent and starts a new join phase or a new discovery phase.

In order to avoid creation of peer loop sequences the new join or discovery phase does not start until all the peer of the S subtree are inhibited to accept join request. In fact every peer knows only the peers that communicate directly with it and can not preview which one try to complete a join with a peer in its sub-tree. In order to avoid deadlock situation at least one children for each peer should be a child one and not a frozen or rejected one.

When S terminates the Join phase sends the new parent message to R that can reallocate the free resources and complete the Join phase with his best frozen child that is not necessarily Q.

3.4. Failure recovery phase

Peer failure leads to a quality degradation of the received services. In fact all the peer in a particular description subtree whose root is the failed peer Q do not receive the descriptions forwarded by Q. We point out that, where separate description paths are allowed, the only effect of a peer failure is a service degradation and not an out of service.

In order to recover the failure and deliver the requested descriptions to the peer in the subtree an alive message is periodically sent by every peer, for example Q, to its own parent node, R. When R receives the alive message, it replies with a aliveToo message. Hence R learns that Q is alive, and Q estimates the current RTT. If R does not receive any alive message in an interval of time which is significantly larger than the alive message sending period, hence the forward of the particular description to Q is interrupted. Q also estimates the average packet jitter of the received descriptions. If there are no data packet arrival in an interval of time which is significantly larger than the estimated jitter an alive message is sent. After the third alive message without reply Q starts a new discovery phase and searches a new parent node; vice versa Q simply waits that some node between its parent and the root performs the recovery phase.
4. Evaluating performances

We evaluate the following figures and traces by simulating sessions between peers localized in 19 Italian universities associated to GARR consortium. In order to estimate the latency between every different university we used the King tool, [4], that allow us to calculate a round trip time probability density function for each couple of selected university. Channel correlation is not considered.

4.1. Relative Delay Penalty

In this paragraph we evaluate the effectiveness of our distributed algorithm. In section 3 we point out that short tree means little delay penalty and low probability of ancestor node failure. In figure 4 the path delay, compared to an ideal situation in which the description source can serve all the other peers with an overlay star topology network, is depicted. We point out that our approach reduce significantly a potentially large delay penalty: the introduced delay are of the same order than the optimal ones. Moreover some peer can experience a lower delay than the optimal one, if the direct path delay are greater than the sum of the delay of the established path. Our simulator also calculate a fixed delay of 0.01 seconds due to the Peer elaboration.

4.2. Failure recovery and Targeted QOS

As explained in P3P different descriptions are delivered through different and independent multicasting trees. With this approach a peer failure does not completely compromise the service delivery in its own subtree. Unfortunately disjoined trees lead to different path delays, but all the descriptions of the same data have to be delivered within the maximum buffer allowed jitter. In figure 5 we depict how often a peer receives n of the 5 delivered descriptions after the peer 1 failure. The effectiveness of the soft-handover strategy is obvious: almost all Peer receive more than 4 description for all the 30 seconds after the failure. We point out that recovery algorithm is already started i.e., the amount of all description data of almost all the peers is greater than the peer 1 amount, hence the perceived quality increases in the following and not depicted seconds. Moreover our channel model does not account channel correlation hence our receiver buffer, that can store 0.3 seconds of data, can be reasonably reduced.

5. Conclusions

In this paper a new, completely distributed overlay multicast platform is proposed. The aim of our overlay network is delivery real time data to a set of peer and allow to a group of participant to join together in single a session without a dedicated infrastructure and without resource wasting. The discovery procedure permit to the participant to know each other without a rendez-vous server i.e., there is the only need of publish or send by mail a small _p3p_ that contains the parameters need to identify of the data source or one of the peers that are just connected to the overlay network. The protocol used by peers try to optimize the tree overlay network and every peer gains a position into the structure that is a function of the below IP network characteristics in term of delay between peer and of the peer available bandwith. Our simulations show that the structure introduced delay are of the same magnitude of the optimal delay, hence no tedious effects are introduced. Moreover a great resilient properties are exploited by introducing a Multiple Description Coding techniques. If each description is delivered by an independent tree the effect of a peer failure is only a temporally quality of service reduction.

References


Abstract—Battery powered wireless sensor networks (WSN) have as main constraint the energy available. The hardware project have to take as objective the power consumption minimization during idle-sleeping sensor states. The software power control and the communication stack implementations have a heavy impact on device power requirement so a carefully design is able to assure longer network lifetime.

In this paper we address the implementation aspects of a WSN geographical routing algorithm [1] in order to investigate the real performances obtainable on a commercially available platform.

Some changes to original design are proposed so as to overcome the problems introduced by sensor devices hardware and operating system limits.

A detailed description of protocol realization is reported and energy measures are made on the devices to characterize the power requirement of the algorithm.

1. Introduction

The applications number involving sensor network are increasing day by day. A lot of new generation commercial devices integrate high data rate radio, and assure compatibility with the new standard 802.15.4. The radio chip installed is often very powerful and more complex than a simple transceiver (see Telos[2] or MicaZ [3] devices) but it does not permit a completely free user defined packet structure. This limit resizes the research field to the high level of communication stack, making difficult to test particular MAC algorithms. Furthermore they operate on 2.4Ghz with possible interference to all the other devices using this band.

On this work EYES sensor devices made from Infineon Technologies have been selected as implementation platform. The main advantage of this devices is represented by the radio transceiver TDA5250 operating on 868Mhz ISM band and produced by Infineon Technologies. The absence of a integrated radio packet control permits the high flexibility to decide each bit sent to the medium, with a little extra overhead for the microcontroller because every byte received has to be immediately processed.

The wireless sensor use the microcontroller MSP430F1161 produced by Texas Instruments (also employed on another new sensors project [2]) that have the interesting capability to switch between five Low Power Mode (LPM) reaching a minimum consumption of about 0.2μA. Some other components presents on the board are dedicated to interface the sensors and to support the USB communication chip.

These qualities combined with the high radio suppleness make the above platform as an ideal choice for the objective put before. More information about the sensor node can be founded on [4].

2. Protocol Description

The multi-hop MAC-Routing protocol implemented uses geographical position and residual energy information to forward the packets to a sink. Radio chips with integrated location engine like CC2431 [5] starts to be available on the market. This new emerging integrated hardware localization solution can overcome the biggest drawback of GeRaF algorithms exploiting all they potentiality.

The main goal of this protocol is the energy consumption minimization; in order to reach this target all the sensors use an aggressive power off strategy with a fixed duty cycle. No synchronization is needed between sensors. The protocol can select hop by hop the best relay in order to reach the network sink.

When a sensor A decides to send a data packet, it switches to ON state and it begins to sense the radio channel for a time $T_{sense}$. This operation is done sampling the RSSI (Received Signal Strength Index) supplied from radio hardware and continually comparing it with a dynamically estimated noise threshold. A packet sent from awake sensors during two reserved time windows permit to reduce the sense time and partially avoid hidden terminal problems.

When $T_{sense}$ time is elapsed without radio channel activity sensing, the sensor sends a Probe packet to recognize the presence of other awake and idle sensors. The time after the Probe packet transmission is implicitly divided on NSlot time slots decided from Probe sender.

Each sensor that receives a Probe and decides to participate to the contention, select a slot number and goes to sleep until its slot starts. The candidate relay sensors wake up during the selected slot and send their Alive packets. It is possible that more than one sensor selects the same slot; on
this case no energy is wasted to resolve the collision, the relay selection procedure uses only the correctly received \textit{Alive} packets. At the end of slots a \textit{SelectRelay} communicate to all neighbors node what is the relay decided from \textit{Probe} sender.

The selected relay remains awake and all the other sensors return to sleep state to save energy as much as possible. Immediately after, the \textit{Data} packet is sent directly to the relay sensor, then \textit{Data} sender switches to the listening state. The selected relay, after the complete reception of \textit{Data} packet, sends immediately the \textit{Probe} packet necessary to next hop. This \textit{Probe} act also as acknowledge for the just terminated hop data transfer.

A back-off mechanism is provided to resolve busy channel sense or transmission errors.

The routing is done using residual energy and a distance index $D_i$ is estimated on possible Relay nodes regarding the destination position. Those information are sent on the \textit{Alive} packets and processed by \textit{Probe} sender in order to take the best possible decision, immediately after the last slot termination the sensor \textit{Probe} sender calculate for each correctly received \textit{Alive} a convenience index as linear combination (with equal weight) of an energy index $E_i$(based on node residual energy), and the hop distance index $D_i$, so it take a decision selecting the relay with minor overall index. Further details about protocol functionalities can be found on [1].

3. Protocol Implementation and Optimization

The code written to implement the algorithm and the main application uses as operating system (OS) a porting of TinyOS [6] for the used eyesIFX hardware platform. This OS integrates on a C language based environment a derived nesC programming language that is able to support a modular composition of the applications with a minimal calculation resource requirement.

The few bit \textit{µ}packet used as collision avoidance mechanism on original protocol design [1] is implemented on a different manner using low level radio functionality. The radio transceiver is put on FSK \textit{Tx} working state for a fixed time, instead of wasting time and energy sending a little packet without information. Two modules \textit{RSSIVccC} and \textit{RSSISeqSampleCMC} are used to implement the Channel sensing functionality and to prevent collision when the radio channel is busy from other transmitters.

The MSP430F1611 energy control is delegated to the \textit{HPLPowerManagementM} that dynamically controls the lowest possible low power mode (LPM) from the five available. Instead the radio chip power off is directly controlled by MAC state machine.

The routing decision mechanism uses the modules \textit{GeoAddrRepositoryC} and \textit{EnergyCounterC} to store and retrieve data about sensor positions and energy consumption. The sensors hardware does not permit a direct measure of current drained from battery and it’s also not possible, to estimate residual energy simply measuring the battery voltage as with a capacitor.

On [7] a technique to balance energy consumption based on battery recovery effect is studied at simulation level but it assume complex battery models as estimation mechanism and is not so easy to implement. It is necessary to report that many factors interferes on battery capacity: temperature, peak current requirement, self discharge; this means that a detailed implementation need to consider all this parameters with a raised computation time. The proposed and implemented solution uses a simple battery life counter with the only target to introduce a fairness on energy consumed from the network sensors. Every time a node starts a transmission or tries to participate to a relay contention this counter is decremented by a fixed value equal to the ‘weight’ of the operation. The 8 MSB of this counter are used as the energy index on \textit{Alive} packets.

The slot timing is carefully designed to overcome radio transceiver switch-on delay. The sensors wake up is anticipated (when possible) by a fixed time in order to have the radio ready to transmit as soon as the slot starts further reducing slots duration.

All the positions information collected from a sensor node are stored on a last recently used table (LRU) implemented on module \textit{GeoAddrRepositoryC}.

In order to distribute complex distance calculus necessary to maximize the hop length, all the distance computation are devolved to the possible relay sensors. The sender sensor has only to compare the difference from its distance to destination value and the possible relay distance, to the maximum radio range. With this optimization during a contention this counter is decremented by a fixed value equal to the main clock is less than few MHz [8]). The square root algorithm implemented is an optimized version for fixed point arithmetic, in order to further reduce computation load.

Due to the impossibility to fix maximum radio range [1], the algorithm implemented obtains the 8 bit position index using as reference the maximum distance covered on a single hop by a packet sent from the current node. On this point the implementation differs from the original algorithm where the radio range is assumed known.

4. Sensor Power consumption measurement

In order to better tune the \textit{EnergyCounterC} module and understand where are possible optimization points, a set of power measures are made on a sensor running the described MAC-Routing protocol.

A constant voltage 3V power supply with an operational amplifier current to voltage converter is used to sense the
sensor current absorption (see figure 1).

A 1 Ω 1% current sense resistor is selected to maintain voltage drop very low and limit measure error. A LeCroy 9400 oscilloscope with 10000 samples recording capability is used with a PC to acquire the current values. The sensor under test uses a data rate of 19200 bps, the NSlot parameter is fixed to 5 and time slot length is selected as 150% of Alive packet duration in order to overcome possible synchronization delay. The TDA5250 radio chip is configured to use the maximum Tx Power and the high LNA (Low Noise Amplifier) sensibility. The figure 2 shows the current curve shape during a packet relay transition recorded from instruments.

At the begin the sensor node is idle and listening the radio channel with a power consumption of about 35mW. At time −0.05 the sensor starts to receive a Probe packet, it is not the sink but it decides to send its Alive packet during the second time slot. The node is so switched to the low power mode available waiting the selected slot. During sleep mode the power consumed is about 4mW.

When the second slot starts the sensor returns to its full working state and the power rises quickly to 37mW. After 10ms the radio state is stable and the Alive transmission can begin requiring about 45.5mW.

When the slots end the sensor returns awake and it receives firstly the SelectRelay and immediately after the Data packet. The peak at about 0.18s is the Probe packet sent to prepare next hop handshake. Now the sensor remains awake to receive Alive packets.

Unfortunately no packets are received and a null SelectRelay is sent to end the communication. Finally the sensor goes to a back-off state and schedules a later re-transmission. It’s interesting to note the negative peak present every time the radio switches from RX to TX state, probably due to the radio chipset used.

The results of these measures shown that when the sensor decides to participate to a communication, the energy consumed (for the slot sleep mechanism) is less than the one used on idle state listening on radio channel. This suggests that, in order to estimate power consumption, the major effect is due to the packet transmission. The packet sending and back-off mechanism is independent from periodic sleep and can heavily influence sensors energy status creating a nonhomogeneous battery residual power distribution on the whole network. The above considerations indicate that the use of a simple transmissions counter can be an interesting first rate approximation of the energy consumed.

5. Network Measurement

To test extensively the presented protocol an area of 10m × 20m is prepared with 45 sensors placed on a rectangular grid of 5×9 points. The Sink is placed on the middle of the shortest grid side.

A PC connected via an USB cable to the Sink collects the packets and through an ad-hoc written Java application [9] it displays and records routing and energy information. To better identify operating network conditions a traffic parameter is defined as:

\[ T = \frac{N_{TOT}}{\lambda \cdot N_{SINK} \cdot S_{DC}} \]

\( N_{TOT} \) is the network sensors number on the coverage area. Packets are generated with an uniform distribution characterized by the parameter \( \lambda \) as mean inter-packet generation time. \( N_{SINK} \) is the number of sinks present on the network. \( S_{DC} \) (Sink drain capability) is an index defined as below:

\[ S_{DC} = \frac{1}{T_{DATA}} \cdot \alpha_{C} \]

\( T_{DATA} \) is the time necessary to receive correctly a packet assuming that no transmission errors happen. \( \alpha_{C} \) is a cor-
retransmission factor that reduces $S_{DC}$ when more than one sink is present on the network.

As whole network configuration a traffic $T = 0.9$ is selected. This selection puts the network on a very high stress condition (maximum allowable traffic is 1). The Duty cycle parameter is fixed to 0.25 with a $Ton = 5s$ and $Toff = 15s$. All the back-off times are randomly generated on sensors with an uniform distribution between 300ms and 5s. A maximum number of tentative is set to 15. When this limit is reached the sensor reset itself and it returns to the IDLE state. During the measure all the sensors generate a new packet only if the MAC is on the IDLE state, otherwise the packet is simply not sent. The figure 3 shows the cumulative probability of packets latency. The latency time is breaded into two parts: the first includes only the first hop, and the second is the mean of the other hops. It is interesting to observe that the minimum delay observed during measures is around 0.5 during first hop and 0.2 on other hops. This difference is due to a protocol optimization that uses the channel sense only during first packet transmission or after a back-off. The results are very good with a high probability to have latency times near to the minimum, also with this high traffic conditions. At the end of measure analyzing data recorded it is important to report, that only a negligible number of packets are dropped on the network ($< 0.2\%$) on this critical condition. However this result can be improved increasing the maximum back-off tentative number.

The figure 4 shows the number of re transmission breaking the results between retransmission for busy channel and retransmission for no Alive reply. On this case the retransmission for no Alive replays are limited on most cases to one, and the main factor is the high traffic that impose an higher re transmission rate for channel busy sense.

6. Conclusion and Future works

An implementation of [1] is presented. A simple and efficient energy estimation module are projected in order to maintain network energy consumption fairness. A distributed computation scheme are realized to reduce single sensor distance square root calculus. Some important optimization are reported in order to reduce protocol energy requirement. Some measure are made on single sensor to characterize protocol power requirement and also some test on whole network are reported. Future works are needed to study an improved algorithm to select different slots number on different network traffic conditions, and the realization of a wider measure on mean whole network energy consumption. More work is also necessary in order to investigate the sensor power requirement during sleep state, and the selection of some solutions to further reduce it.

References

[5] Chipcon CC2431 (SoC) solution for ZigBee/IEEE 802.15.4 wireless sensor networking http://www.chipcon.com
Forced Synchronization of Class I and Class II Neurons with both Electrical and Chemical Synapses

Eri Ioka† and Hiroyuki Kitajima‡

Faculty of Engineering, Kagawa University
2217-20 Hayashi, Takamatsu, Kagawa, Japan
Phone:+81-87-2226 FAX:+81-87-2262
†Email: s05g451@stmail.eng.kagawa-u.ac.jp
‡Email: kitaji@eng.kagawa-u.ac.jp

Abstract—Neurons produce repetitive firing patterns by an external stimulus. By the difference of the onsets of the firing, neurons are classified into two types, class I and class II. We compare forced synchronization in class I and class II neurons with electrical (gap junction) and chemical synapses. We obtain that the class II neurons have wider parameter regions of in-phase forced synchronization than the class I neurons have.

1. Introduction

Neurons produce repetitive firing patterns by an external stimulus. Many researchers verify that oscillation dynamics of neural activity and its synchronization play an important role in the models of information processing in the brain. The oscillation mechanisms of neuron models are classified into two by their bifurcations: class I and class II[1], [2]. The oscillation of former is associated with the saddle-node bifurcation which shows zero frequency at the bifurcation point. On the other hand, the latter is related to the Andronov-Hopf bifurcation exhibiting a finite frequency at the bifurcation point.

We have considered forced synchronization in electrically coupled two Morris-Lecar(ML) neurons with both class I and class II excitabilities. Izhikevich and Ermentrout showed that class II neurons are easy to achieve synchronization [3], [4]. However, by the detailed bifurcation analysis, Tsumi et al. indicated that class I neurons also have various synchronized states and a lot of their bifurcations[5]. However, these results are only for mutual synchronization, and the results of forced synchronization are not reported as far as we know.

We have studied the differences of forced synchronization between class I and class II neurons with gap junction connection. We have showed that class II neurons have wider parameter regions of forced synchronization than those of class I neurons [6]. The bifurcation structure for class II neurons are complicated; there are period-doubling(subcritical) bifurcations.

Currently, many researchers propose that the inhibitory neurons play an important role in the synchronous firing, and verify that inhibitory neurons have both gap junctions and chemical synapses in the neocortex[7]-[9].

In this paper, we consider the two ML neurons coupled by both synapses and compare forced synchronization between class I and class II neurons. We find that the class II neurons have wider parameter regions of in-phase synchronization than the class I neurons have.

2. System Equation

We analysis the coupled two ML neurons with chemical and electrical synapses described as follows:

\[ C_M \frac{dV_i}{dt} = -g_i(V_i - V_i) - g_{Ca}M_{\infty}(V_i - V_{Ca}) - g_K N_i(V_i - V_K) + g_{gap}(V_{gap} - V_i) + g_{syn}s_i(V_{syn} - V_i) \] (1)

\[ \frac{dN_i}{dt} = \frac{N_{\infty} - N_i}{\tau_{N_i}} \] (2)

\[ \frac{ds_i}{dt} = \frac{s_{\infty} - s_i}{\tau_{s_i}} \] (3)

\[ M_{\infty} = 0.5[1 + \tanh((V_i - V_a)/V_b)] \] (4)

\[ N_{\infty} = 0.5[1 + \tanh((V_i - V_c)/V_d)] \] (5)

\[ \tau_{N_i} = 1/[\phi \cosh((V_i - V_c)/2V_d)] \] (6)

\[ s_{\infty} = 0.5[1 + \tanh((V_i + V_{th})/V_{slope})] \] (7)

where \( V \) is the membrane potential of the neuron, \( N \) is the activation variable for \( K^+ \) . \( \tau_N \), \( M_{\infty} \) and \( N_{\infty} \) are functions of \( V \). \( s_i \) is the synaptic channel, and \( s_{\infty} \) is synaptic channels at steady state[10]. \( I_{ext} = I + I_m \sin(\omega t) \), \( g_{gap} \) and \( g_{syn} \) are the maximal conductance

971
of the electrical and the chemical synapses current, respectively. The ML model can be changed between class I and class II excitabilities by the value of the parameter $V_c$. The critical point is about $V_c = 4.6$.[11]. In this paper, we choose $V_c = 12$ and $V_c = 2$ for the class I and the class II neurons, respectively. We fix the DC current $I$ as 50 and 55 for the class I and the class II neurons, respectively, to obtain an equivalent oscillation frequency in the single neuron. The other parameter values are fixed as follows:

Table 1: The values of parameter in Eq. (1)-(7).

\[
g_{Ca} = 4.0 \text{mS/cm}^2, \quad g_{K} = 8 \text{mS/cm}^2, \quad g_{L} = 2 \text{mS/cm}^2, \\
V_{Ca} = 120 \text{mV}, \quad V_k = -80 \text{mV}, \quad V_h = 18 \text{mV}, \\
V_l = -1.2 \text{mV}, \quad V_I = 17.4 \text{mV}, \quad V_c = 1 \text{mV}, \\
V_{th} = 0 \text{mV}, \quad \tau_e = 1 \text{msec}, \\
C_m = 20 \text{mF/cm}^2, \quad \phi = 1/15 \text{sec}^{-1}.\]

We fix the values of $g_{gap}$ and $g_{syn}$ as 0.01 and 1.0, respectively, to obtain anti-phase synchronization when $I_m = 0$. The aim of this study is to investigate the phase transition from anti-phase to in-phase as the amplitude $I_m$ of common external forces is increased. The reversal potential of the synaptic coupling conductance $V_{syn} = -84$, and the resting potential (RP) of the ML system is about -65. Because $V_{syn}$ smaller than RP, that neuron model is inhibitory.

3. Result

We analyze the class I and class II neurons, assuming that the amplitude of the external force $I_m$ is $0 \leq I_m \leq 20$, and the angular frequency $\omega$ is $0 \leq \omega \leq 0.23$. Time scale is changed to $\omega t$ for fixing the period of the external force as $2\pi$. Bifurcation sets in the two parameter plane ($\omega, I_m$) are used.

Figures 1(a) and 1(b) are bifurcation diagrams for the class I and class II neurons, respectively, obtained by the brute-force method and Kawakami’s method[12]. In these figures horizontal and vertical axes indicate the angular frequency ($\omega$) and the amplitude ($I_m$) of the external alternative current, respectively. In these figures, the solid curves indicate the saddle-node bifurcations, thick dashed ones express the period-doubling bifurcations, and tiny dashed ones symbolize the pitchfork bifurcations. The blue and green shaded areas illustrate the region of in-phase and non-in-phase synchronous firing, respectively. The red and white ones indicate the region of in-phase and anti-phase quasi-periodic solutions, respectively. In Figs. 1(a) and 1(b), a natural frequency is 0.054 and 0.057 for the class I and class II neurons. The blue curves indicate bifurcations of sub-harmonic oscillation of order two which is an anti-phase solution. In Fig. 1(b), we observe that the period-doubling and pitchfork bifurcations occur for the fundamental oscillation. Moreover, Figs. 1(a) and 1(b) show that the class II neurons have wider in-phase synchronous regions than class I.

Figure 2(a) and 2(b) are bifurcation diagrams for both classes neurons with only electrical synapses. The white shaded areas indicate the region of quasi-periodic solutions. In Figs. 2(a) and 2(b), a natural frequency is 0.077 and 0.079 for the class I and class II neurons. The blue curves indicate bifurcations of sub-harmonic oscillation of order two which is an anti-phase solution. Comparing Figs. 1(b) and 2(b), the bifurcation structure of fundamental oscillation is similar. However, the case of chemical synapse coupling is more complicated than the electrical one’s.

Figure 3(a) and 3(b) are one-parameter bifurcation diagrams, when we fix $I_m = 15$ and move $\omega t$ along the arrow $l$ in Figs. 1(a) and 1(b) between 0.01 and 0.11. The green and red points are the Poincaré mapping for the membrane potential of neuron 1 and neuron 2, respectively. We find that the class II neurons have wider parameter regions of forced synchronization than the class I neurons, because the class II neurons have more periodic points than the class I neurons have. Therefore, the class II neurons easily achieve forced synchronization compared with the class I neurons.

Figure 4(a) and 4(b) are waveforms of membrane potentials $V_1$ and $V_2$ at the points(a) in Figs. 3(a) and 3(b), respectively. Then, we fix $I_m = 15$ and $\omega = 0.105$. In Figs. 4(a) and 4(b), the green and red curves show the membrane potentials of the neuron 1 and neuron 2, respectively. These figures shows that the neuron 1 and neuron 2 synchronize in anti-phase for both class I and class II neurons.

4. Conclusion

we study the bifurcation structure of the class I and class II neurons with gap junction and chemical synapses. We compare them, and find that only class II neurons have period-doubling and pitchfork bifurcations. And, we find that the class II neurons have wider parameter regions of in-phase forced synchronization than the class I neurons have.
Acknowledgments

The authors would like to thank Dr. K. Tsumoto of Aihara Complexity Modelling Project, ERATO, JST for his helpful comments.

References


Figure 1: Bifurcation diagram. Solid and dashed curves indicate the saddle-node and the period-doubling bifurcation, respectively. We observe stable in-phase synchronization states in the blue and red regions; in each figure black and blue curves indicate the bifurcation structure of fundamental oscillation for in-phase solution and sub-harmonic oscillation of order two for anti-phase solution, respectively.
Figure 2: Bifurcation diagram. The meaning of curves and regions except white regions is the same as Fig. 1.

Figure 3: One-parameter bifurcation diagram on the line of \( I_m = 15.0 \) in Figs. 1(a) and 1(b). The green and red dots are the Poincaré mapping at each time \( t = 2\pi j \) \( (j = 1, 2, \cdots, 10000) \) for the membrane potential of neuron 1 and neuron 2, respectively.

Figure 4: The waveforms at the points(a) \( (I_m = 15.0, \omega = 0.105) \) in Figs. 3(a) and 3(b). The green and red curves show the membrane potentials of the neuron 1 and neuron 2, respectively.
Critical Temperature Sensor Based on Spiking Neuron Models
Gessyca Maria Tovar, Tetsuya Hirose, Tetsuya Asai, and Yoshihito Amemiya

Graduate School of Information Science & Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo, 060-0814, Japan
Email: gessyca@sapiens-ei.eng.hokudai.ac.jp, hirose@sapiens-ei.eng.hokudai.ac.jp, asai@sapiens-ei.eng.hokudai.ac.jp, amemiya@sapiens-ei.eng.hokudai.ac.jp

Abstract
We developed a subthreshold CMOS circuit whose dynamical behavior, e.g., oscillatory or stationary behaviors, changes at a given threshold temperature with the aim of the development of low-power and compact temperature switch devices on monolithic ICs. The threshold temperature is set to a desired value by adjusting an external bias voltage. The circuit consists of two pMOS differential pairs, capacitors, a current reference circuit with low-temperature dependence, and two off-chip resistors.

1. Introduction
Temperature control is fundamental in various electronic systems. There are several passive and active sensors for measuring system temperature, including thermocouples, resistive temperature sensors, thermistors, and silicon temperature sensors [1]. Among present temperature sensors, a thermistor that has positive temperature coefficients (PTC) is widely used because it exhibits a sharp increase of its resistance at a specific temperature. It is suitable for implementation in temperature control systems that make decisions, such as over-temperature shutdown, turn-on/off cooling fan, temperature compensation, or general-purpose temperature monitor.

Temperature characteristics of PTC thermistors made from ceramics, such as sintered metal oxides, are based on its material characteristics and its relative proportions [1], [2]. Electrical resistivities of PTCs are shifted by a phenomenon called leaching, which occurs during the soldering process. In addition to the resistance shift, leaching causes degradation of the solder-electrode and electrode-semiconductor bond, which may result in thermistors having greatly reduced stability and reliability. Therefore, it is difficult to build a PTC thermistor on monolithic ICs, which prevents us from developing on-chip temperature control systems.

We developed an analog circuit that has a sharp transition characteristic similar to that of PTC thermistors. Key features of the circuit include low-power subthreshold operation of MOSFETs, extended range of threshold temperature, and small packages.

2. The Model
Temperature increase causes a regular and reproducible increase in the frequency of pacemaker potential in most _Aplysia_ and _Helix_ excitable neurons [3]. The Br-type neuron, located in the right parietal ganglion of _Helix pomatia_, only shows its characteristic bursting activity between 12 and 30°C. Outside this range, the burst pattern disappears and the action potentials become regular. In other words, excitable neurons can be used as a sensor to determine the temperature range in a natural environment.

There are many excitable neuron models, but only a few of them were previously implemented on CMOS LSIs, e.g., silicon neurons that emulate cortical pyramidal neurons [4], FitzHugh-Nagumo neurons with negative resistive circuits [5], artificial neuron circuits based on by-products of conventional digital circuits [6]-[8], and ultralow-power subthreshold neuron circuits [9]. We developed a model based on the Wilson-Cowan system [10] because of ease of theoretical analysis and subthreshold CMOS implementation.

Figure 1 shows the temperature sensor operation model. The temperature sensor consists of a nonlinear oscillator that changes its state e.g., oscillatory or stationary states when it receives an external perturbation (temperature).

The dynamics of the temperature sensor can be expressed as:

\[
\dot{u} = -u + \frac{e^{u/A}}{e^{u/A} + e^{v/A}} 
\] 

(1)
Fig. 2: $u$ and $v$ nullclines with vector field direction

and

$$\dot{v} = -v + \frac{e^{u/A}}{e^{u/A} + e^{\theta/A}}$$

(2)

where $\tau$ is the time constant, $A$ is a constant proportional to temperature and $\theta$ is a variable. The second term of the r.h.s. of Eq. (1) is the sigmoid function, a mathematical function that produces an S-shaped (sigmoid) curve.

To analyze the system operation, we must calculate the nullclines, which are curves in the phase space where $\dot{u}$ and $\dot{v}$ are zero. Nullclines divide the phase space into four regions, and in each region, the vector field follows a specific direction. On the $u$ nullcline, the direction of the vector is vertical, while on the $v$ nullcline, it is horizontal. Figure 2 shows $u$ and $v$ nullclines and the vector field in each region.

The trajectory also depends on $\tau$, which modifies the velocity field of $u$. In Eq. (1), if $\tau$ is large, the value of $u$ decreases, and for small $\tau$, $u$ increases. Figures 3-a and b show trajectories when $\tau=1$ and $\tau<<1$. The trajectory of $u$ is much faster than that of $v$, so only close to the $u$ nullcline direction of movement other than horizontal is possible.

To explain the operation of the system, let us suppose that $\theta$ is set at certain value, where the critical temperature ($T_c$) is $27^\circ C$. The critical temperature represents the threshold temperature we want to measure.

When $\theta$ changes, the $v$ nullcline changes its position in the phase space to a point where the system will be stable as long as the temperature is higher than $T_c$ because the system is unstable only when fixed point exists in a negative resistive region of the $u$ nullcline. The fixed point represents the intersection point of the $u$ and $v$ nullclines. At this point, the trajectory stops because the vector field is zero. On the other hand, when the temperature is below $T_c$, the nullcline and the fixed point change. The trajectory does not pass through the point, and the system starts oscillating. Figures 4-a and b show the trajectories when the system is stable and oscillatory.

The local minimum $(u_-,v_-)$ and local maximum $(u_+,v_+)$ where the nullclines intersect are given by:

$$u_\pm = \frac{1 \pm \sqrt{1 - 4A}}{2}$$

(3)

and

$$v_\pm = u_\pm + A\ln\left(\frac{1}{u_\pm} - 1\right)$$

(4)

Figure 5 shows the local minimum and maximum represented in the phase space.

Combining the local minimum and maximum with Eq. (2), we determine a relationship between $\theta$ and temperature, given by:
3. The circuit

Figure 6 shows our critical temperature sensor circuit, which consists of two pMOS differential pairs (M₁-M₂ and M₃-M₄), two capacitors (C₁ and C₂), a current reference circuit [11], and two off-chip metal-film resistors (g) with low-temperature dependence.

Differential pairs subthreshold currents \( I_1 \) and \( I_2 \) are given by:

\[
I_1 = I_b \frac{e^{\kappa u/v_T}}{e^{\kappa u/v_T} + e^{\kappa v/v_T}}
\]

and

\[
I_2 = I_b \frac{e^{\kappa v/v_T}}{e^{\kappa u/v_T} + e^{\kappa v/v_T}},
\]

where \( \kappa \) is subthreshold slope, \( v_T \) is thermal voltage (\( v_T = KT/q \)), \( K \) is Boltzmann’s constant, \( T \) is temperature, \( q \) is elementary charge, \( c_1 \) and \( c_2 \) are the capacitance, \( \theta \) is bias voltage.

The circuit changes its dynamical behavior between stable and oscillatory at a given critical temperature, which is set by adjusting the bias voltage \( \theta \).

Numerical simulations were conducted by setting \( c_1 \) and \( c_2 \) at 0.1 and 10 pF, respectively, \( g \) at 1 nS, and reference current \( I_b \) at 1 nA. Figure 7 shows the nullclines and trajectory of the circuit with \( \theta \) set at 0.5 V and \( T \) at 27°C; the system is in an oscillatory state.

Setting \( \theta \) until the system changes its state, we established a numerical relation between \( \theta \) and \( T_c \): \( \theta \) for \( u \) local minimums and \( \theta \) for \( v \) local maximums.

Figure 8 shows the relation between \( \theta \) and \( T_c \). When \( \theta \) is used to set \( T_c \), the system is stable at temperatures higher than \( T_c \); while when \( \theta \) is used, the system is stable when the temperature is lower than \( T_c \) and oscillatory when it is higher than \( T_c \).

Figure 9 shows the numerical nullclines and trajectory using \( \theta \), \( T \) is higher than \( T_c \); and Fig. 10 shows the numerical nullclines and trajectory using \( \theta \), \( T \) is lower than \( T_c \).

We used SPICE simulation to determine circuit operation. We simulated circuit nullclines and changed...
4. Summary

We developed a subthreshold CMOS circuit whose dynamical behavior, e.g., oscillatory or stationary behaviors, changes at a given threshold temperature. Threshold temperature was set to a desired value by adjusting bias voltage. The circuit’s operation was investigated through theoretical analysis and extensive numerical simulations.

5. References


Abstract—This paper addresses a discrete-time consensus scheme for dynamically changing multi-agent systems with nonlinear performance functions. The information flows among agents are represented as a time-varying weighted digraph. We analyze the global stability of the system by algebraic graph theory and the mean value theorem. This method allows us to deduce the conditions on the edge weights and on the graph topologies which guarantee a consensus in the group of agents.

1. Introduction

Recently, there has been much attention to cooperative control for multi-agent systems. This is due to a vast range of applications such as a vehicle formation of unmanned air vehicles (UAVs), swarming and flocking behavior of mobile robots, and sensor networks [1]. In cooperative control for multi-agent systems, each agent is required to act in concert with each other, and to make decisions independently at the same time. Moreover, the group of agents is expected to achieve a team objective even in dynamic and unpredictable environments.

Therefore, designing a proper control law has become a central issue for cooperative control. Ren et al. study the information consensus problem under unreliable information exchanges with dynamically changing interaction topologies [2]. They discuss linear discrete-time and continuous-time consensus schemes by algebraic graph theory. In [3], weighted gradient methods for resource allocation is considered. This focuses on a distributed resource allocation problem with a time-invariant undirected graph.

In this paper, we present the design framework for cooperative control systems which achieves an agreement among a group of agents in a distributed fashion. We consider a nonlinear and discrete-time consensus algorithm for dynamic multi-agent systems whose information flows are modeled as a time-varying weighted digraph. Our purpose here is to equalize the performance of agents by imposing conditions on edge weights of the interaction graph and on the graph topologies.

This paper is organized as follows. We begin by reviewing mathematical properties of the graph theory and the matrix theory in Section 2. In Section 3, we consider a consensus protocol for a decentralized coordination. A stability analysis for the proposed protocol is discussed in Section 4. In Section 5, we provide simulation experiments to demonstrate the validity of our theoretical results. Finally, we state concluding remarks in Section 6.

2. Preliminaries

One approach to a consensus problem for multi-agent systems is the algebraic graph theory which observes graph properties based on linear algebra and its applications. In this section, we summarize mathematical preliminaries from the graph theory and the matrix theory [2], [4], [5], [6].

2.1. Graph

The interaction topology among n agents is modeled as a weighted digraph $G(V, E)$ with a node set $V = \{v_i | i \in I\}$ and an edge set $E \subseteq V \times V$, where $I = \{1, 2, \ldots, n\}$. Each node $v_i$ represents agent $i$ and each directed edge $(v_i, v_j) \in E$ is interpreted as an unidirectional communication from agent $i$ to agent $j$. A directed tree is the directed graph whose nodes except the root have exactly one parent. A spanning tree of a digraph is the directed tree formed by graph edges that connect all the nodes of the graph. A directed graph has a spanning tree if it contains a spanning tree as a subgraph.

Let $G_i(V, E_i)$ $(i = 1, 2, \ldots, M)$ be the finite possible interaction graphs with a common node set $V$. The union of graphs $G = \bigcup_{i=1}^{M} G_i$ is the digraph with the common node set $V$ and the union of the edge set $\bigcup_{i=1}^{M} E_i$.

2.2. SIA Matrix

Let $M_n(R)$ be a set of all $n \times n$ real matrices. A matrix $P = \{p_{ij}\} \in M_n(R)$ is called nonnegative, denoted by $P \geq 0$, if $p_{ij} \geq 0$ for all $i, j \in I$. A (row) stochastic matrix $P \in M_n(R)$ is the nonnegative matrix which satisfies $p_{ij} \geq 0$ for all $i, j \in I$, and $\sum_{j=1}^{n} p_{ij} = 1$. A stochastic matrix $P$ is said to be indecomposable and aperiodic (SIA) if $\lim_{m \to \infty} P^m = \mathbf{1} \mathbf{y}^T$, where $\mathbf{1}$ denotes the $n \times 1$ column vector whose components are all one and $y$ is an $n \times 1$ column vector [5]. Two nonnegative matrices $A, B \in M_n(R)$ are said to be the same type if their zero elements and nonzero elements are in the same locations [5].

Lemma 1 ([2], [6]) Let $S_P = \{P_1, P_2, \ldots, P_l\}$ be a set of stochastic matrices with positive diagonal entries. If the graph associated with $P_i$ $(i = 1, 2, \ldots, l)$ has a spanning tree, then $P_i$ is SIA. Moreover, if the union of the graphs of matrices $P_i$ $(i = 1, 2, \ldots, l)$ has a spanning tree, then the
matrix product $\prod_{i=1}^{t} P_i$ is SIA.

For a stochastic matrix $P = [p_{ij}] \in M_n(\mathbb{R})$ $(i, j \in I)$, let

$$\lambda(P) = 1 - \min_{i,j \in I} \sum_{i=1}^{n} \min(p_{i,j}, p_{j,i}).$$  \hspace{1cm} (1)

Lemma 2 ([2], [5], [6]) Let $S = \{S_1, S_2, \ldots, S_k\}$ be the finite set of SIA matrices with the property that, for each sequence $S_i, S_{i+1}, \ldots, S_k$ of positive length, the matrix product $S_i S_{i+1} \cdots S_k$ is SIA. Then for each infinite sequence $S_i, S_{i+1}, \ldots$, there exists a column vector $v$ such that

$$\lim_{t \to \infty} S_i S_{i+1} \cdots S_k = v^T.$$  \hspace{1cm} (2)

In addition, in the case that $S$ is an infinite set, $\lambda(S) < 1$, where $S = S_1 S_2 \cdots S_k$ and $N_t$ is defined as the number of different types of all $n \times n$ SIA matrices. Furthermore, if there exists a constant $0 \leq \gamma < 1$ satisfying $\lambda(S) \leq \gamma$, then Eq. (2) also holds.

3. Consensus Protocol

Consensus protocol is the action policy for agents which determines their own actions, interaction rules with other agents, and reactions against an external environment. In this section, we propose a nonlinear discrete-time consensus protocol to execute cooperative control in a distributed manner.

Let $x_i[k] \in \mathbb{R}$ be an information state of agent $i$ $(i \in I)$ at step $k$ $(k = 0, 1, 2, \ldots)$. A performance corresponding to an information state $x_i$ is characterized by a performance function $\phi_i : \mathbb{R} \to \mathbb{R}$ and a performance value of agent $i$ at step $k$ is denoted by $\phi_i[k] = \phi_i(x_i[k])$. We assume that each performance function $\phi_i(x_i)$ is strictly increasing and of class $C^1$. Let $a_{ij}[k]$ be a positive time-varying weight on edge $(v_i, v_j) \in E$ and $g_{ij}[k]$ be the time-varying Boolean variable which represents the presence of edge $(v_i, v_j)$:

$$g_{ij}[k] = \begin{cases} 1, & \text{if agent } i \text{ receives information from agent } j \text{ at step } k, \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, we assume that all agents can obtain their own information at all steps, i.e. $g_{ij}[k] = 1$ for any $k$.

We consider the following nonlinear discrete-time consensus protocol:

$$x_i[k+1] = x_i[k] + \sum_{j=1}^{n} a_{ij}[k] g_{ij}[k] (\phi_j[k] - \phi_i[k]),$$  \hspace{1cm} (3)

$$\phi_i[k] = \phi_i(x_i[k]).$$  \hspace{1cm} (4)

Equation (3) indicates that each agent updates its information state in order to narrow the performance gap between the agent and its neighbors. For example, if there exists an information path from agent $j$ to agent $i$ and the performance of agent $j$ is better (resp. worse) than the one of agent $i$, agent $i$ upgrades (resp. degrades) its own information state according to the performance difference between agent $i$ and agent $j$. Equation (4) means that a performance of each agent is evaluated by its information state.

Suppose $x_i[k+1] \neq x_i[k]$. By the mean value theorem, there is $x'_i[k]$ such that

$$\frac{\phi_i(x_i[k+1]) - \phi_i(x_i[k])}{x_i[k+1] - x_i[k]} = \phi'_i(x'_i[k]),$$  \hspace{1cm} (5)

where $\min(x_i[k], x_i[k+1]) < x'_i[k] < \max(x_i[k], x_i[k+1])$.

Let us define $\beta_i[k] (i \in I)$ as

$$\beta_i[k] = \begin{cases} \phi_i'(x'_i[k]), & x_i[k+1] \neq x_i[k], \\ \phi_i'(x_i[k]), & x_i[k+1] = x_i[k]. \end{cases}$$  \hspace{1cm} (6)

From Eqs. (5) and (6), if $x_i[k+1] \neq x_i[k]$, we have

$$\phi_i[k+1] - \phi_i[k] = \beta_i[k] (x_i[k+1] - x_i[k]).$$  \hspace{1cm} (7)

Notice that Eq. (7) also holds for $x_i[k+1] = x_i[k]$. Define matrices $A[k]$ and $B[k]$ as

$$A[k] = \begin{bmatrix} - \sum_{j=1}^{n} a_{ij}[k] c_{ij}[k] & a_{ij}[k] c_{ij}[k] & \cdots & a_{ij}[k] c_{ij}[k] \\ a_{ij}[k] c_{ij}[k] & - \sum_{j=1}^{n} a_{ij}[k] c_{ij}[k] & \cdots & a_{ij}[k] c_{ij}[k] \\ \vdots & \vdots & \ddots & \vdots \\ a_{ij}[k] c_{ij}[k] & a_{ij}[k] c_{ij}[k] & \cdots & - \sum_{j=1}^{n} a_{ij}[k] c_{ij}[k] \end{bmatrix},$$

and $B[k] = \text{diag}[\beta_1[k], \beta_2[k], \ldots, \beta_n[k]]$. Then the consensus protocols (3) and (4) are rewritten as follows:

$$x_i[k+1] = x_i[k] + A[k] \phi_i[k],$$  \hspace{1cm} (8)

$$\phi_i[k+1] = \phi_i[k] + B[k] (x_i[k+1] - x_i[k]),$$  \hspace{1cm} (9)

where $x_i[k] = [x_i[k], x_2[k], \ldots, x_n[k]]^T$ and $\phi_i[k] = [\phi_i[k], \phi_2[k], \ldots, \phi_n[k]]^T$.

We say that consensus about the performance of agents is achieved asymptotically if, for any initial information state $x_i[0] (i \in I)$,

$$|\phi_i[k] - \phi_j[k]| \to 0, \text{ as } k \to \infty,$$  \hspace{1cm} (10)

for any $i, j \in I$.

We can apply the proposed consensus protocols (3) and (4) to a dynamic resource allocation problem. In this case, an information state of agent $i$ can be interpreted as an allocated resource to agent $i$.

4. Stability Analysis

In this section, we investigate how to guarantee an agreement in a team of agents. To this end, we deduce the conditions about the edge weights of interaction topologies. We also impose the condition on the interaction topologies to ensure that each agent communicates with other agents sufficiently often.

Lemma 3 Let $d^n_i[k]$ be the number of incoming edges to node $v_i (i \in I)$. Diagonal entries of a matrix $W[k]$ are all positive if any time-varying weight $a_{ij}[k]$ on edge $(v_i, v_j) \in E$ satisfies

$$0 < a_{ij}[k] < \frac{1}{d^n_i[k] \beta_{ij}^{\text{max}}},$$  \hspace{1cm} (11)

for $d^n_i[k] \neq 0$, where $\beta_{ij}^{\text{max}} = \max(a_{ij}[k])$.  


Proof: Let \( d_i^0[k] \neq 0 (i \in I) \). By Eq. (11), we obtain
\[
0 < \alpha_i[k] < \frac{1}{d_i^0[k] \beta_i^0} \leq \frac{1}{d_i^0[k] \beta_i[k]},
\]
which implies
\[
1 - \beta_i[k] d_i^0[k] r_{ij}[k] > 0.
\]
(13)
Note that Eq. (13) also holds for \( d_i^{in}[k] = 0 \). Let \( \alpha_i^{max} = \max_{j=1,...,w_t} \{ \alpha_i[k] | g_{ij}[k] = 1 \} \). Then we have
\[
1 - \beta_i[k] d_i^{in}[k] \alpha_i^{max}[k] \\
\geq 1 - \beta_i[k] d_i^0[k] \alpha_i^{max}[k] > 0.
\]
There is an interesting property for the consensus protocols (3) and (4) by setting additional conditions as described in the next lemma.

Lemma 4 Suppose each performance function \( \phi_i(x_i) \) \( (i \in I) \) satisfies \( \phi_i(x_i) \geq 0 \) for \( x_i \geq 0 \), and \( \phi_i(0) = 0 \). Let \( x_i[0] > 0 \). Then an information state \( x_i[k] \) is positive for each \( k = 0, 1, 2, \ldots \) with the consensus protocols (3) and (4) if any time-varying weight \( \alpha_{ij}[k] \) on edge \((v_j, v_i) \in E \) satisfies
\[
0 < \alpha_{ij}[k] < \frac{1}{d_i^{in}[k] \beta_i^{max}}
\]
for \( d_i^{in}[k] \neq 0 \), where \( \beta_i^{max} = \max_j \beta_i[j] \).

Proof: Let \( x_i[k] > 0 \) \( (i \in I) \) for any \( k = 0, 1, 2, \ldots \). By Eq. (4), we have
\[
\phi_{ij} = \frac{\phi_i(x_i[k])}{x_i[k]}. \]
(15)
Define \( \gamma_i[k] \) \( (i \in I) \) by \( \gamma_i[k] = \frac{\phi_i(x_i[k])}{x_i[k]} \) and a diagonal matrix \( C[k] \) by \( C[k] = \text{diag} \{ \gamma_1[k], \gamma_2[k], \ldots, \gamma_n[k] \} \). Note that \( \gamma_i[k] \geq 0 \) for any \( i \in I \) from the assumptions \( x_i[k] > 0 \) and \( \phi_i(x_i) \geq 0 \) for any \( i \in I \). Then we can rewrite Eq. (15) as
\[
\phi_i[k] = C[k] x_i[k].
\]
(16)
From Eqs. (8) and (16), we obtain
\[
x[k+1] = x[k] + A[k] C[k] x[k] = (I + A[k] C[k]) x[k]. \]
(17)
Now suppose \( \gamma_i^{max} = \max_i \gamma_i[k] \), then
\[
\gamma_i^{max} \leq \frac{\phi_i(x_i[k+1]) - \phi_i(x_i[k])}{x_i[k+1] - x_i[k]} \\
\leq \max_i \phi_i(x_i) = \beta_i^{max}.
\]
Thus, if Eq. (14) holds for \( d_i^{in}[k] \neq 0 \), we have
\[
0 < \alpha_{ij}[k] \leq \frac{1}{d_i^{in}[k] \beta_i^{max}}
\]
which implies
\[
1 - \gamma_i[k] \sum_{j=1,...,w_t} \alpha_{ij}[k] g_{ij}[k] \\
\geq 1 - \gamma_i^{max} d_i^{in}[k] \alpha_i^{max}[k] > 0,
\]
and hence \( I + A[k] C[k] \) is a nonnegative matrix with positive diagonal entries. On the other hand, if \( d_i^{in}[k] = 0 \), it is clear that \( x_i[k+1] = x_i[k] \). Therefore, \( x_i[k+1] \) is also positive.

Considering the above argument and the initial condition \( x_i[0] > 0 \), we conclude that \( x_i[k] \) is positive for all \( k = 0, 1, 2, \ldots \) under the conditions of Lemma 4.

By extending the work of Ren et al. [6], we have the following theorem which guarantees a consensus among a group of agents.

Theorem 1 Let \( x_i[0] > 0 \) \( (i \in I) \). All performance values \( \phi_i[k] \) starting from any initial state \( \phi_i(x_i[0]) \) converge asymptotically to an equilibrium point using the consensus protocols (3) and (4) if there exist \( \epsilon > 0 \) \( (\epsilon \in R) \) and a time-varying weight \( \alpha_{ij}[k] \) on edge \((v_j, v_i) \in E \) such that
\[
\epsilon < \alpha_{ij}[k] < \frac{1}{d_i^{in}[k] \beta_i^{max}}
\]
for \( d_i^{in}[k] \neq 0 \), and there exists an infinite sequence \( 0 = k_0 < k_1 < k_2 < \cdots \) such that
(i) the time interval \([0, \infty) \) is divided into nonoverlapping subintervals \([k_0, k_1) \cup [k_1, k_2) \cup [k_2, k_3) \cup \cdots \)
(ii) each subinterval \([k_m, k_{m+1}) \) \( (m = 0, 1, 2, \ldots) \) is uniformly bounded and the union of the interaction graphs across each subinterval has a spanning tree.

Proof: From the condition (20), \( \alpha_{ij}[k] \) is lower and upper bounded. By Lemma 3, each \( V[k] \) \( (k = 0, 1, 2, \ldots) \) is a stochastic matrix with positive diagonal entries and all nonzero entries of \( W[k] \) are lower bounded. We know that there exists an infinite sequence \( 0 = k_0 < k_1 < k_2 < \cdots \) with conditions (i) and (ii). Define \( W^{(m)} \) \( (m = 0, 1, 2, \ldots) \) as the product of matrices \( W[k] \) over the \((m+1)-st \) subinterval \([k_m, k_{m+1}) \). Then each \( W^{(m)} \) is SIA since the graph associated with \( W^{(m)} \) has a spanning tree. In addition, any product of \( W^{(m)} \) is also SIA. A matrix \( W^{(m)} \) only has finite types since possible locations of nonzero and zero entries in the \( n \times n \) matrix \( W^{(m)} \) are finite. Note also that nonzero entries of each \( W^{(m)} \) are lower bounded. Now let
\[
\overline{W} = W^{(1)} W^{(2)} \cdots W^{(s_0)},
\]
(21)
where \( N_t \) is the number of different types of all \( n \times n \) SIA matrices \( W^{(m)} \) \( (m = 0, 1, 2, \ldots) \). Noting that we consider the time interval \([0, \infty) \), \( \overline{W} = W^{(m_0)} W^{(m_0+1)} \cdots \) is an infinite set. Hence, we obtain \( \lambda(\overline{W}) < 1 \) from Lemma 2. From the definition of \( N_t \), the types of \( \overline{W} \) are finite and nonzero entries of \( \overline{W} \) are lower bounded. Thus, there exists \( 0 < q_i < 1 \) \( (i = 1, 2, \ldots, \overline{w}_t) \) such that \( \lambda(\overline{W}) \leq q_i \) for each \( \overline{W} \), where \( \overline{w}_t \) is the number of different types of \( \overline{W} \). Moreover,
by setting \( q = \max\{q_1, q_2, \ldots, q_w \} \), there exists \( 0 \leq q < 1 \) such that \( \lambda(W) \leq q \) for any \( W \). From Lemma 2, we have
\[
\lim_{j \to \infty} W^{(m)}_j W^{(m-1)} \cdots W^{(m)} = 1 \mu^T,
\]
(22)
where \( \mu \) is a nonnegative \( n \times 1 \) vector and \( \mu \neq 0 \). This immediately follows that
\[
\lim_{k \to \infty} \phi[k] \phi[k-1] \cdots \phi[1] = c \mathbf{1},
\]
(23)
where \( c \in \mathbb{R} \) is a constant value.

The condition (20) in Theorem 1 is suitable for distributed multi-agent systems since this condition only requires the maximum derivative value of the performance function \( \phi(x_i) \), and the number of incoming information links to agent \( i \). Moreover, the existence of the sequence \( 0 = k_0 < k_1 < k_2 < \cdots \) satisfying conditions (i) and (ii) does not require a full connection and allows the failure or the suspension of the communication among agents if each agent communicates frequently enough.

5. Simulation Results

We illustrate our method by the simulation below. We consider a group of 5 agents with dynamically changing interaction topologies whose union in each infinitely consecutive subinterval has a spanning tree. The evolution of the interaction topologies from Step 1 to Step 4 is shown in Figure 1.

![Figure 1: The evolution of interaction topologies from Step 1 to Step 4 with 5 agents.](image)

In this simulation, we choose the initial values of the information state as \( x_1[0] = 0.20, x_2[0] = 0.12, x_3[0] = 0.18, x_4[0] = 0.24 \), and \( x_5[0] = 0.26 \). The performance functions are given as follows:

\[
\phi_1(x_1) = \begin{cases} \frac{1}{2} + \frac{1}{2} \sin \left( \pi \left( x_1 - \frac{1}{2} \right) \right), & 0 \leq x_1 < 1, \\ \frac{1}{2}, & 1 \leq x_1, \end{cases}
\]

\[
\phi_2(x_2) = \begin{cases} 1 + \sin \left( \frac{x_2 - 1}{2} \right), & 0 \leq x_2 < 1, \\ \frac{x_2 + 1}{2}, & 1 \leq x_2, \end{cases}
\]

\[
\phi_3(x_3) = x_3, \quad 0 \leq x_3,
\]

\[
\phi_4(x_4) = \begin{cases} \sin \left( \frac{x_4}{2} \right), & 0 \leq x_4 < 1, \\ \frac{x_4}{2}, & 1 \leq x_4, \end{cases}
\]

\[
\phi_5(x_5) = \begin{cases} x_5, & 0 \leq x_5 < 1, \\ 2x_5 - 1, & 1 \leq x_5. \end{cases}
\]

Therefore, we have \( \beta_1^{\text{max}} = \frac{\pi}{2}, \beta_2^{\text{max}} = \frac{\pi}{2}, \beta_3^{\text{max}} = 1, \beta_4^{\text{max}} = \beta_5^{\text{max}} = 2 \).

We set the weight \( a_{ij}[k] \) on edge \((v_j, v_i) \in E\) as
\[
a_{ij}[k] = 0.9 \times \frac{1}{d_i^m[k] \beta_i^{\text{max}}} \tag{24}
\]
for \( d_i^m[k] \neq 0 \). Note that these assumptions satisfy the conditions of Lemma 4 and Theorem 1.

![Figure 2: The convergence process of performance values \( \phi[k] \) and the corresponding changes of information states \( x_i[k] \) for \( i = 1, 2, \ldots, 5 \).](image)

Figure 2 shows the changes of the performance values \( \phi[k] \) and the information states \( x_i[k] \) for \( i = 1, 2, \ldots, 5 \). We can see that consensus among the group of agents is reached and that information states are positive at all steps.

6. Conclusions

In this paper, we considered a nonlinear discrete-time consensus framework for dynamic multi-agent systems. This method enabled us to derive sufficient conditions for a consensus among agents about their performances. The conditions of Theorem 1 were well-suited to designing a distributed coordination algorithm since each agent does not need to know the global information. We also provided a numerical example to illustrate our theory by simulation. However, the condition about an edge weight (20) in Theorem 1 may be conservative considering that the edge weight is required to be positive and bounded by the maximum value of \( \phi'(x_i) \) for each agent. Relaxing this condition will be our future work.

References


Learning Partial Synchronization Regimes with Imposed Qualitative Behavior on an Array of Chua’s Oscillators

D´aniel Hillier††, Serkan G¨unel†, Johan A.K. Suykens§ and Joos Vandewalle†

†ESAT/SCD-SISTA, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
‡Jedlik Laboratory, P´eter P´azm´any Catholic University, Pr´ater u. 50/a, 1083 Budapest, Hungary
Email: hillier@digitus.itk.ppke.hu, serkan.gunel,johan.suykens,joos.vandewalle@esat.kuleuven.be

Abstract—We demonstrate that using a numerical optimization framework it is possible to explore and analyze the information processing capabilities of non-linear oscillator networks. It will be shown how a properly formulated cost function can be used to achieve synchronization and at the same time impose qualitative behavior on an array of chaotic oscillators. Our approach aims to add an application motivated aspect to existing results that so far focused on conditions for synchronization. To demonstrate the potential of the optimization framework we show on a simple case study that network configurations corresponding to partial synchronization regimes can be learned. This can also be done with imposing various qualitative behavior on individual oscillators.

1. Introduction

Research on synchronization in coupled chaotic oscillators was so far mainly focusing on deriving conditions for synchronization. In [6] global symmetries of a network of identical, diffusively coupled oscillators were used to classify linear invariant manifolds corresponding to partial synchronization regimes and to analyze their global asymptotic stability. In [2] the link between graph topology and stability of global synchronization was clarified. In [1] it was shown that stability of partial synchronization regimes depends on the individual oscillator dynamics and on the choice of state variables used in the coupling. However the existence of partially synchronized clusters depends only on the network topology.

These findings are of primary importance on the way to unravel the rules that govern complex dynamics witnessed in networks of coupled oscillators. The advantage is that these rules reflect the most fundamental mathematical properties of coupled oscillators independent of individual dynamics of oscillators and from initial conditions. However it is not clear how these results can be applied with affordable computational complexity to larger networks or to networks with a structure that is not very simple. Real-world coupled dynamical systems are much more complex than those used in the above mentioned studies.

This paper presents a case study showing how global optimization methods can be used to reveal rules of synchronization potentially playing a role in information processing.

2. Statement of the problem

We consider an array composed of nonlinear oscillators:

\[
\dot{x}_{rc} = F(x_{rc}; A_{rc}) + \sum_{k,j \neq c} G_{jk} r_{jk} x_{rk} \quad r, c = 1, \ldots, M \in \mathbb{N}_0
\]

where \( x_{rc} \in \mathbb{R}^d \) represents the state vector of the \( r, c \)-th oscillator; \( F : \mathbb{R}^d \times \mathbb{R}^p \rightarrow \mathbb{R}^d \) is a nonlinear function depending on parameters \( A_{rc} \in \mathbb{R}^{d \times p} \) that define the cells; \( G \in \mathbb{R}^{M \times M} \) is a weight matrix and \( P = (p_{ij}) \in \mathbb{R}^{d \times d} \) with \( p_{ij} \in [0, 1] \) determines which state variables couple the oscillators and \( S_p(r,c) \) defines a fixed neighborhood topology with neighborhood radius \( \rho \). We investigated how a network with specific properties can be learned, e.g. by solving the optimization problem:

\[
\min_{A, G} U((x_{01}(t_0), x_{01}(t_1), \ldots, x_{01}(t_n)) (2)
\]

where \( U(\cdot) \) denotes the cost function and \( X_{st} = (x_{st}(t_0), x_{st}(t_1), \ldots, x_{st}(t_n)) : \mathbb{R} \rightarrow \mathbb{R}^{Md} \) is a stationary solution of (1) for a given initial condition and \( t_0 < t_1 < \ldots < t_n < \infty \). Additional inequality constraints may apply to \( A \) and \( G. U(\cdot) \) is constructed in such a way that dynamical properties of individual cells or cell populations fulfill the desired requirements. These requirements may include different types of synchronization and stability criteria, desired Lyapunov spectrum or embedding dimension of the attractor, etc. We assume that \( F_{rc} \) can provide a rich enough set of dynamical behaviors.

One fundamental issue to be solved is to find network configurations that make a given set of cells in the array synchronize with additional requirements on qualitative behavior. An important advantage of the proposed approach compared to other studies is that the only requirement on the vector field defining individual oscillators is that they permit the solution to exist and being unique. In this study we used oscillators with continuous vectors fields. On the other hand, in some cases this liberty may result in an optimization problem that is very hard to solve if it is possible at all.
3. Global Optimization Framework

Inspired by e.g. [7] where it was shown that difficult tasks can be learned for dynamical systems, we conceived a novel approach on the way to investigate the signal and image processing ability of a network defined above. The global optimization framework is generic, i.e. no assumption was made about network size or topology. The system-level diagram of the framework is shown on Figure 1.

The first task is to define a proper cost function. The most straightforward cost function uses only the time evolution of the network. Theoretical knowledge can also be incorporated, e.g. inspired by [2] we can prescribe a specific information traffic distribution for links between cells to be synchronized and for links to the rest of the array. This way theoretical results can be transformed into system design laws.

Definition of the training set means that we explicitly prescribe which cells are expected to be synchronized. Alternatively we can only prescribe a number of cells that have to be synchronized. The latter approach is easier to formulate and can be used as a first test to check whether the evolution of the network confirms that the cost function was properly formulated.

Prescribing specific cell configurations to be synchronized is a more tedious task - the number of possibilities grows exponentially even if we drop repeating configurations due to the rotational symmetry of a rectangular array. This extra complexity however pays off abundantly since it can reveal the capabilities and limitations of the network in well defined signal and image processing tasks. The choice of optimization algorithm is an intricate task. Since integrating the network for a given parameter set can last long, a key factor here is to choose algorithms that require a low number of function evaluations to find the global optimum.

In [1] examples were shown indicating that partial (phase) synchronization regimes persist even with 100% of parameter mismatch. Although we lack generic theoretical knowledge about how parameter variation affects synchronization error, such examples - together with results shown later in this paper - indicate that cost surfaces minimizing synchronization error are smooth with several local minima. Therefore a candidate optimization algorithm must perform global search first.

In [8] a new global optimization algorithm called Coupled Simulated Annealing (CSA) is introduced. In CSA the annealing temperatures of several Simulated Annealing processes are interconnected in order to improve performance in convergence speed and to increase the probability of exploring all basins of attraction in a given number of cost function evaluations. In addition, the coupling also allows the control of the variance of the acceptance probabilities by the acceptance temperature. The sensitivity of the acceptance temperature on its initial value is therefore reduced. In addition, variance control guides optimization to quasi-optimal runs, i.e. it balances global search done at high temperatures with gradually switching to local search with the temperature cooling. The number of cost function evaluations per individual processes decreases exponentially when the number of optimizers is increased linearly. This also makes CSA a good candidate since interactions between solutions decrease the number of cost function evaluations to reach a given energy threshold. Once a solution is found in the framework, its sensitivity to variations in initial conditions and parameter variations have to be tested.

![Figure 1: Overview of the investigated global optimization framework that teaches a network of chaotic oscillators to match qualitative behavior requirements formulated in the cost function.](image)

4. Case study: Partial synchronization with imposed dynamics on a coupled array composed of Chua’s oscillators

To demonstrate the potential of the optimization framework we applied it to investigate for rules related to the onset and stability of partial synchronization regimes.

4.1. Cost function for synchronization

Defining a cost function for synchronization is a very intricate task since there is no widely accepted measure of synchronization. During our study we investigated for a method that is

- able to capture different forms of synchronization (complete-, phase-, lag-, generalized)
- provides the measure in a way that does not need any human interpretation, i.e. can be embedded in the cost function
Currently we analyzed complete synchronization of a prescribed configuration of cells. This aim can be reached in a computationally cheap manner by minimizing

\[ U_{\text{std}} = \sum_{i=1}^{n} \sum_{j=1}^{d} \exp \left( \frac{1}{\sum_{k \in \Pi G} \left( x'_i(t_j) - \bar{\tau}_S(t_j) \right)^2} \right) \]  

where \( x'_i(t_j) \in \Pi S \) represents a number of snapshots taken from the time evolution of cells to be synchronized, \( x'_i(t_j) \in \Pi R \) represents those that are not to be synchronized. \( \Pi S, \Pi R \subset [1, ..., M] \times [1, ..., M] \) denote cell populations with \( N_S \) standing for the total number of cells to be synchronized and \( N_R \) for the number of cells to be desynchronized and \( \Pi S \cap \Pi R = \emptyset \). \( \bar{\tau}_S(t_j) \) is the spatial average for each time instance of the target population of cells to be synchronized and \( \bar{\tau}_R(t_j) \) is the average for the rest of the cells. Briefly, for a set of time instances we require the spatial standard deviation of the target population to vanish whereas the ratio of spatial standard deviations between the target population and the rest is maximized.

### 4.1.1. Cost function imposing qualitative behavior

In this study we consider chaotic, limit cycle and equilibrium point qualitative behavior. Lyapunov exponents are usually used to measure qualitative properties of a dynamical system. The cost function used to impose any of these modes on either the population required to be in sync or on the rest is:

\[ U_{\text{sys}}(\Omega, G) = \sum_{j \in \Theta_e} e^{-\lambda_{\text{max}}(x_j; A, G)} + \sum_{j \in \Theta_R} e^{\lambda_{\text{max}}(x_j; A, G)} + \sum_{j \in \Theta_s} |\lambda_{\text{max}}(x_j; A, G)| \]  

where \( \lambda_{\text{max}} \) is the maximum Lyapunov exponent estimated from the time series of an oscillator (for a fast implementation see [4]). Cells are also partitioned into mutually exclusive sets according to the qualitative behavior we want to impose on them after initial transients lapsed during the evolution of the array. Chaotic behavior is imposed on cells denoted by \( \Theta_e \) convergence to one of the equilibria is imposed on cells denoted by \( \Theta_s \) and limit cycle behavior on cells denoted by \( \Theta_l \).

### 5. Simulation results

We performed learning for 2x2, 3x3 and 7x7 on arrays of Chua’s oscillators with hyperbolic tangent non-linearity, with eight-connected local topology, zero-flux boundary conditions and simulated in Matlab with evolution times large enough to analyze the long-term stationary behavior of the array (\( T = 200 \)). According to [5], the dynamics of a cell is given by:

\[ \frac{d}{dt} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -a & -a & 0 \end{pmatrix} \begin{pmatrix} x^1 \\ x^2 \\ x^3 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -af(x^1) \end{pmatrix}. \]  

Variable \( x^1 \) was coupled to \( x^3 \), i.e. \( P = 0 \ 0 \ 0 ; \ 0 \ 0 \ 0 ; \ 1 \ 0 \ 0 \) in (1). Both the coupling matrix and the matrix of bifurcation parameters (\( a \) in (5)) of each cell were included in the learning process. On Figure 2 we show illustrative results for the 2x2 case where the target population to be synchronized was the bottom-left and upper-right cells. CSA with 16 individual optimizers and 200 population iterations ran for 20 minutes using ode45 with relative tolerance 10^{-3} on a Linux machine powered by an AMD Opteron 250. Similar qualitative effects can be observed for 3x3 and 7x7 arrays. An important result of this study is that we showed that it is possible to learn partial synchronization regimes on arrays having different qualitative behavior. Figure 2 shows two examples for partial synchronization: chaotic dynamics emerges when limit cycle and convergent oscillators are coupled, but the same cells can synchronize also when coupling makes chaotic oscillators become limit-cycle oscillators.

### 6. Discussion and conclusions

In [3] Liu et al. claim that when coupling is added to a network of chaotic oscillators with double- or multi-scroll attractor, Lyapunov exponents being zero in the uncoupled system become positive. They suggest that this rule is general, however, the example shown on Figure 2 indicates that this is not always true since we were able to learn configurations where the opposite effect was happening. This example demonstrates, that the proposed approach can be a useful tool to test hypotheses made on coupled networks of chaotic oscillators. Rules (either theoretical or empirical) believed as generic can be transformed into a cost function. Then a global optimization process can maximize the cost function, i.e. it looks for solutions that violate the rule formulated as a cost function. Finding no such solution can be a strong evidence supporting the validity of the original rule. Confirming generic (theoretical) rules via global optimization might seem improper. However, deriving sound theoretical results is extremely difficult when dealing with coupled networks of chaotic oscillators. Therefore the method showed in this paper can provide valuable feedback. This study showed that by using a global optimization framework it is possible to get insight into the complex interactions of coupled chaotic oscillators. Our method can be useful in guiding investigations and confirming results from theoretical analysis.
Acknowledgments

The authors would like to thank Samuel Xavier de Souza for fruitful discussions and Tamás Roska for his support. Research supported by the Research Council KUL: GOA-AMBioRICS, CoE EF/05/006 Optimization in Engineering, Flemish Government: FWO: G.0211.05 (Nonlinear Systems), G.0226.06 (Cooperative Systems) Belgian Federal Science Policy Office IUAP P5/22 and The Scientific and Technological Research Council of Turkey (TUBITAK).

References


Figure 2: Trajectories showing the stationary behavior from time $T$ to $2T$ of two qualitative behavior types of Chua’s oscillator arrays. For both types, cells (2,1) and (1,2) were learned to synchronize. For a,b,c) positive maximum Lyapunov exponents ($\lambda_{\text{max}}$) were imposed on all cells. For c,d,f): zero $\lambda_{\text{max}}$ were imposed on all cells. On a)-d) continuous line denotes synchronizing cells, dotted line denotes desynchronization. On a,c) trajectories of the array without coupling are shown. b) and d) illustrate the behavior when coupling is added to array a) and c) respectively. e,f) Synchronization errors shown as $\text{norm}(X_{\text{REF}})$ versus $\text{norm}(X_c)$ between a chosen reference cell and all other cells. Errors of array b) and d) are shown on e) and f) respectively. The coupling and bifurcation parameters $G_{\text{c,c}}$ and $A_{\text{c,cr}}$ are shown below each oscillator.
Global Noise Estimation Based on Tensor Product Expansion with Absolute Error

Akitoshi ITAI†, Hiroshi YASUKAWA†, Ichi TAKUMI†† and Masayasu HATA‡

†Aichi Prefectural University, 1522-3 Kumabari, Nagakute, Aichi, 480-1198 JAPAN
‡†Nagoya Institute of Technology, Gokiso-cho, Syowa-ku, Nagoya-shi, Aichi, 466-8555 JAPAN
¶Collage of Engineering, Chubu University, 1200 Matsumoto-cho, Kasugai-shi, Aichi, 487-8501 JAPAN
E-mail:†a_itai@cis.aichi-pu.ac.jp, yasukawa@ist.aichi-pu.ac.jp
††takumi@ics.nitech.ac.jp, ‡hata@cs.chubu.ac.jp

Abstract—This paper proposes a novel signal estimation method that uses a tensor product expansion (TPE). When a bivariable function, which is expressed by two-dimensional matrix, is subjected to conventional TPE, two single variable functions (SVFs) are calculated by minimizing the mean square error between the input vector and its outer product. TPE is useful for feature extraction and signal compression, however, it is difficult to separate global noise from other signals. This paper shows that global noise, which is observed in almost all input signals, can be estimated by using a tensor product expansion where absolute error is used as the error function.

1. Introduction

It is well known that multivariate analysis is useful for extracting the features of bivariable functions. The major method of extracting the characteristics of second-order statistics like principal component analysis (PCA) produces some global feature of various data. PCA is useful for feature extraction, however, it is difficult to separate global and local features raised by different sources. TPE[1] can approximate an m-variable function as the outer product using m SVFs. This technique is applied to nonlinear system identification[2], 3-D image processing[3], and achieve a substantial results, respectively. The tensor product composed of two SVFs is calculated by minimizing a mean square error between the input vector and its outer product. It is difficult to separate global noise from other signals since TPE is the same as PCA if m = 2. In this paper, we propose a method of a signal estimation that uses TPE and applies absolute error to separate a local signal from global noise; the former is observed in just few signals while the latter is seen in most signals. Simulations show that proposed method is useful for estimating global noise.

2. Tensor product expansion

The tensor product of m-variable function is obtained by the product of m SVFs that minimizes the mean square error. This calculation is called tensor product expansion. In most time-series analyses of observed signals (i.e. sounds, biological signal), the input vector is a bivariable function expressed by two-dimensional matrix. TPE with a bivariable function is written as:

$$\sum_{l_1=0}^{q_1} \sum_{l_2=0}^{q_2} (h(l_1, l_2) - f_1(l_1)f_2(l_2))^2$$  \hspace{1cm} (1)

where \(l_1 \text{ and } l_2\) are variables that show the row and the column, \(h(l_1, l_2)\) is a bivariable input vector, \(f_1(l_1)\) and \(f_2(l_2)\) are SVFs, \(q_1\) and \(q_2\) are the length of row and column. Bivariable TPE is the same as singular value decomposition, hence, it can be considered that \(f_1(l_1), f_2(l_2)\) are the first principal component and its eigenvector, respectively. In this paper, \(l_1\) indicates the index of the time courses, \(l_2\) shows the index of the observation site.

3. Tensor product expansion using absolute error (TPE-AE)

TPE can extract features of the bivariable function (input vector), however, it is difficult to extract anomalous or global noise. In order to separate these signals using TPE, we propose the absolute error as a new evaluation function (TPE-AE). Using absolute error, the tensor of a bivariable function is given as:

$$\sum_{l_1=0}^{q_1} \sum_{l_2=0}^{q_2} |h(l_1, l_2) - f_1(l_1)f_2(l_2)|.$$  \hspace{1cm} (2)

Tensor product \(f_1(l_1)f_2(l_2)\) in Eq.2 approximates only the AC components, since the equation does not have an additional term, which approximates a DC component. The following equation is proposed to obtain the AC and DC components of \(h(l_1, l_2)\):

$$\sum_{l_1=0}^{q_1} \sum_{l_2=0}^{q_2} |h(l_1, l_2) - (f_1(l_1)f_2(l_2) + f_3(l_2))|$$  \hspace{1cm} (3)
where \( f_1(l_1)f_2(l_2) \) and \( f_3(l_2) \) approximate the AC and DC components, respectively. Assume that \( s_a(l_1) \) and \( s_b(l_1) \) are the source signal of global noise and a local signal, respectively, and \( A_{l_2} \) and \( B_{l_2} \) are a mixed matrix and DC component, respectively. A steady signal that includes only global noise is obtained by the following expression:

\[
A_{l_2}s_a(l_1) + B_{l_2}
\]  

(4)

and an unsteady signal that includes global noise and a local signal is given by:

\[
A_{l_2}s_a(l_1) + s_b(l_1) + B_{l_2}.
\]  

(5)

In this section, we show that the optimal \( f_1, f_2, f_3 \), which minimize Eq. 3, is given by a linear expression of global noise. If the time course \( f_1(l_1) \) is approximated as \( s_a(l_1) \) in the following expression, global noise can be separated from an observed signal.

\[
f_1(l_1) = A_s a(l_1) + B.
\]  

(6)

If a steady and an unsteady signal are observed in input vector \( h(l_1, l_2) \), we have

\[
h(l_1, 0) = A_0s_a(l_1) + B_0
\]  

(7)

\[
h(l_1, 1) = A_1s_a(l_1) + s_b(l_1) + B_1.
\]  

(8)

Substituting Eq.7,8 into Eq.3, we get

\[
\begin{align*}
\sum_{l_1}^{q_1} \left[ A_0s_a(l_1) + B_0 - (f_1(l_1)f_2(0) + f_3(0)) \right] \\
+ \sum_{l_1}^{q_1} \left[ A_1s_a(l_1) + s_b(l_1) + B_1 - (f_1(l_1)f_2(l_1) + f_3(l_1)) \right] \\
= A_0X(l_1) + A_1X(l_1) + s_b(l_1) + s_a(l_1)\]  

(9)

where \( q_1 \) is a length of time course, first (second) term is the absolute error of the steady (unsteady) signal. Here, we assume \( A_i \) and \( B_i \) can be estimated by \( f_2(l_2), f_3(l_2) \), respectively. Eq.9 is expressed as:

\[
\sum_{l_1}^{q_1} \left[ |A_0X(l_1)| + |A_1X(l_1) + s_b(l_1)| \right] + \sum_{l_1}^{q_1} \left[ |A_0X(l_1)| + |A_1X(l_1) + s_b(l_1)| \right] = 0
\]  

(10)

To obtain Eq.10 we used the fact that \( X(l_1) = s_a(l_1) - f_1(l_1) \). Eq.10 has a \( \sum \), however, it is only necessary to estimate optimal \( X(l_1) \) minimizing Eq.10 in each \( l_1 \). Therefore, we have:

\[
\sum_{l_1}^{q_1} \left[ |A_0X(l_1)| + |A_1X(l_1) + s_b(l_1)| \right] = 0
\]  

(11)

Eq.11 includes two absolute terms. Optimal \( X(l_1) \), which minimize Eq.11, is calculated in the following 3 cases.

**Case 1 and 2.** \( A_0 = A_1 \) and \( A_0 < A_1 \)

From \( A_0 = A_1 \) and Eq.11, optimal \( X(l_1) \) in Eq.11 is given by the following expression:

\[
0 \leq X(l_1) \leq s_b(l_1) \\
0 \leq f_1(l_1) \leq s_a(l_1) + s_b(l_1)/A_1
\]  

(12)

Optimal \( X(l_1) \) in the case of \( A_0 < A_1 \) is given as:

\[
\begin{align*}
X(l_1) &= s_0(l_1)/A_1 \\
f_1(l_1) &= s_a(l_1) + s_b(l_1)/A_1
\end{align*}
\]  

(13)

\( f_1 \) in Eq.12.13 includes a local signal, so global noise is not estimated from \( h(l_1, l_2) \)

**Case 3.** \( A_0 > A_1 \)

Optimal \( X(l_1) \) in the case of \( A_0 > A_1 \) is given as:

\[
\begin{align*}
X(l_1) &= 0 \\
f_1(l_1) &= s_a(l_1)
\end{align*}
\]  

(14)

Eq.14 means that \( f_1(l_1) \) is obtained by \( s_a(l_1) \) in the case of \( A_0 > A_1 \). Thus, it is shown that global noise can be estimated by using TPE-AE.

In actual analyses, the number of signals is more than 2. Let the steady and unsteady signal vector be \( x \) and \( y \), respectively. If multiple steady and unsteady signals are observed, we have

\[
\begin{align*}
h(l_1, x) &= A_x s_a(l_1) + B_x \\
h(l_1, y) &= A_y s_a(l_1) + s_b(l_1) + B_y
\end{align*}
\]  

(15)

(16)

Similarly, optimal \( X(l_1) \), which minimizes Eq.11, is obtained by the following expressions.

\[
\begin{align*}
\sum_x A_x &= \sum_y A_y : s_a(l_1) \leq f_1(l_1) \leq s_a(l_1) + s_b(l_1)/A_y \\
\sum_x A_x &> \sum_y A_y : f_1(l_1) = s_a(l_1) + s_b(l_1)/A_y \\
\sum_x A_x &< \sum_y A_y : f_1(l_1) = s_a(l_1)
\end{align*}
\]  

(17)

Eq.17 shows that TPE-AE can estimate global noise \( s_a(l_1) \) that satisfies the following conditions.

- Mixed matrix \( A_i \) and DC component \( B_i \) can be estimated.
- Mixed matrix of steady and unsteady signals \( A_x, A_y \) satisfies \( \sum A_x > \sum A_y \).

Thus, a local signal is separated from an unsteady signal since global noise is estimated as a tensor product.

### 4. Monte Carlo simulation

It is difficult to obtain optimal SVF \( f_1(l_1), f_2(l_2), f_3(l_2) \) using TPE-AE since the evaluation function Eq.3 has a nonlinear term. In this paper, we use a relaxation method based on a Monte Carlo simulation (MCS) to solve TPE-AE. MCS produces optimal solutions by extensive trials using random numbers. Using MCS, optimal SVF, which minimize Eq.3 is obtained by the following steps:
Table 1: Conditions of the artificial signal

<table>
<thead>
<tr>
<th></th>
<th>Global noise</th>
<th>Block pulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>(\sin(2\pi/576)/3.0+N(0.8,0.05^2))</td>
<td>0.3</td>
</tr>
<tr>
<td>B</td>
<td>(\sin(2\pi/576)/3.0+N(1.2,0.05^2))</td>
<td>0.3</td>
</tr>
<tr>
<td>C</td>
<td>(\sin(2\pi/576)/4.0+N(1.0,0.05^2))</td>
<td>0.6</td>
</tr>
<tr>
<td>D</td>
<td>(\sin(2\pi/576)/5.0+N(0.8,0.05^2))</td>
<td>0.6</td>
</tr>
<tr>
<td>E</td>
<td>(\sin(2\pi/576)/5.0+N(1.2,0.05^2))</td>
<td>0.6</td>
</tr>
</tbody>
</table>

: \(N(\mu, \sigma^2)\) means Gaussian noise \(\mu\) is mean, \(\sigma^2\) is variance.

Table 2: Conditions of the input vector and TPE-AE

<table>
<thead>
<tr>
<th>Sampling number ((q_1))</th>
<th>5760</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector number ((q_2))</td>
<td>5</td>
</tr>
<tr>
<td>The number of just sine waves</td>
<td>3</td>
</tr>
<tr>
<td>The number of vector including block pulse</td>
<td>2(B,D)</td>
</tr>
<tr>
<td>Update times</td>
<td>1000</td>
</tr>
</tbody>
</table>

STEP

1. Initialize \(f_1(l_1), f_2(l_2), f_3(l_2)\) using a small random number.
2. \(f_{new1}(l_1)=f_1(l_1)+\text{rand}(l_1)\).
3. In \(l_1, \sum_{l_2} |h(l_1, l_2) - (f_1(l_1)f_2(l_2) + f_3(l_2))| > \sum_{l_2} |h(l_1, l_2) - (f_{new1}(l_1)f_2(l_2) + f_3(l_2))|\) is satisfied, so \(f_1(l_1) = f_{new1}(l_1)\).
4. Repeat step 2 and 3, 10 times.
5. \(f_{new2}(l_2)=f_2(l_2)+\text{rand}(l_2)\).
6. In \(l_2, \sum_{l_1} |h(l_1, l_2) - (f_1(l_1)f_2(l_2) + f_3(l_2))| > \sum_{l_1} |h(l_1, l_2) - (f_1(l_1)f_{new2}(l_2) + f_3(l_2))|\) is satisfied, then \(f_2(l_2) = f_{new2}(l_2)\).
7. Repeat step 5 and 6, 10 times.
8. \(f_{new3}(l_2)=f_3(l_2)+\text{rand}(l_2)\).
9. In \(l_2, \sum_{l_1} |h(l_1, l_2) - (f_1(l_1)f_2(l_2) + f_3(l_2))| > \sum_{l_1} |h(l_1, l_2) - (f_1(l_1)f_2(l_2) + f_{new3}(l_2))|\) is satisfied, then \(f_3(l_2) = f_{new3}(l_2)\).
10. Repeat step 8 and 9, 10 times.

where \(\text{rand}()\) is white Gaussian noise based on normal distribution \(N(0,1)\). The termination criteria should be decided for above iterative algorithm. According to empirical result, almost TPE is updated enough by 1000 update. Therefore, this algorithm is repeated at 1000 times.

5. Simulation

5.1. Known functions

The simulation results of TPE-AE are shown to demonstrate its effectiveness for global noise estimation. First, a simple artificial signal based on known functions is subjected to TPE-AE. The artificial signal consists of a global noise component and a local signal, which are generated by a sine wave and a block pulse train, respectively. Table.1 lists the conditions of the input vector \(h(l_1, l_2)\) shown in Fig.1, where \(l_1\) indicates the index of the time courses, \(l_2\) is the index of the artificial signal A-E (0 to 4). The vertical axis shows the amplitude of an artificial signal, the horizontal axis indicates the cycle of sine wave \((l_1/576)\). In the input vector, 2 signals include a local signal in cycles 7 to 10.

of the input vector \(h(l_1, l_2)\) shown in Fig.1, where \(l_1\) indicates the index of the time courses, \(l_2\) is the index of the artificial signal A-E (0 to 4). The vertical axis shows the amplitude of an artificial signal, the horizontal axis indicates the cycle of sine wave \((l_1/576)\). In the input vector, 2 signals include a local signal in cycles 7 to 10.

The tensor product of conventional TPE is shown in Fig.2, proposed TPE-AE is shown in Fig.3 and the result of subtracting Fig.3 from Fig.1 is shown in Fig.4. Fig.2 shows the combination of the sine wave and block pulse. This means that separating global noise from a local signal is difficult with TPE. Fig.3 shows that the sine wave is estimated by the proposed method, but Gaussian noise is also extracted. From Fig.4, a local signal can be detected by subtracting a tensor product from input vector. Hence, it is shown that global noise and a local signal can be separated by using TPE-AE.

5.2. Global noise reduction in ELF EM wave

We show the effectiveness of TPE-AE in assessing actual observation signals. Attention is being placed on the electromagnetic (EM) waves that radiate from
Table 3: Conditions of the input vector and TPE-AE

<table>
<thead>
<tr>
<th>Sampling number($q_1$)</th>
<th>1152</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vector number($q_2$)</td>
<td>7</td>
</tr>
<tr>
<td>The number of just global noise</td>
<td>6</td>
</tr>
<tr>
<td>The number of vector including local radiation</td>
<td>1</td>
</tr>
<tr>
<td>Update times</td>
<td>1000</td>
</tr>
</tbody>
</table>

the earth’s crust in advance of earthquakes and volcanic activity[4]. Such EM waves observed in the extremely low frequency band include global noise created by lightning radiation from the tropics. This noise has a daily trend and similarity in all stations. If global noise in the EM waves can be estimated, a local signal from the earth’s crust and lightning in the near field can be detected adequately. The input vector for applying TPE-AE is composed of 7 signals captured on July 5-6, 2003 by 7 antennas. 576 data points were collected per day for each component. Table.2 shows the conditions of the input vector. The length of time course is 1152, the number of signals is 7 where 1 signal includes local radiation. Therefore, $l_1$ indicates the index of the time course, $l_2$ is the index of the observation signal 0 to 6. The two observed signals shown in Fig.5 include a daily trend and local radiation. The vertical axis shows the EM level (pT/$\sqrt{\text{Hz}}$), the horizontal axis indicates the day ($l_1/576$).

The TPE result shown in Fig.6 includes global noise and local radiation. On the other hand, the proposed method can estimate the daily global noise from Fig.7. The local radiation observed at 0.7 and 1.6 days can be detected by subtracting Fig.7 from Fig.5 (Fig.8). These results show that TPE-AE is effective in estimating global noise in ELF EM waves.

6. Conclusion

In this paper, we proposed a signal estimation method based on tensor product expansion that uses absolute error. We introduce the separability condition for global noise estimation. Simulation results have shown the effectiveness of the proposed method for the global noise observed most signals. Other simulation results have shown that the proposed method can reduce the daily trend observed in ELF EM waves. However, these simulations do not show the performance for the complex signal separation, since simple input vectors are applied to simulations. A remaining problem is to reduce Gaussian noise and to apply the proposal to other observation signals.

Acknowledgment

This research was supported in part by JSPS for the Grant-in-Aid for Scientific Research (A)17206042.

References


Abstract – The paper deals with the simulation techniques of switching circuits typically employed in power electronics. The focus is on the transient analysis and on its typical issues. A first comparison reports the simulation results obtained utilizing a general purpose simulation software (PSpice) and those obtained by means of the co-simulation PSpice-Simulink. In the second comparison, the simulation results obtained utilizing a simulation software devoted to the analysis of power electronic circuits (PSIM) and those obtained by means of the co-simulation PSIM-Simulink are examined. The last approach has shown the best computational performances.

1. Introduction
   Power electronic circuits can be analyzed by general-purpose simulators such as PSpice [1] and good results are obtained when externally driven switches are considered. Nevertheless, when a circuit employing internally driven switches has to be studied, general purpose simulators show great problems of convergence. In most cases the main effect is a very long simulation time due to the high iterations number and in some cases the simulation is just stopped, because the time-step necessary to continue the simulation is too small. These problems depend on both the switch model (usually a controlled resistor varying in a large range) and on the complexity of the control implementation, for which there are not dedicated libraries.

   Considering that control libraries exist in the Matlab’s toolbox Simulink, a good solution for simulating efficiently the switching circuits could be the co-simulation PSpice-Simulink. In this manner the power block could be implemented in PSpice environment, while the control block could be implemented in Simulink one. The main difficulty of the co-simulation is to allow the two simulator to interact each other. This aspect is guaranteed by an interface tool, called SLPS [2], between the MathWorks’ MATLAB/Simulink system simulator and the PSpice A/D electric circuit simulator.

   If the under test system is not complex, e.g. a single-phase converter, the simulation results are good enough; but if a three phase inverter with the voltage control and the current control has to be analyzed, both of the simulation techniques show difficulties [3]-[4]-[5]-[6].

   Then, for complex systems of power electronic it needs to utilize dedicated software such as PSIM. Even in this case the interaction with the toolbox Simulink is possible by means a software module, called SimCoupler.

   The paper is organized as follow. Section II describes the electric system under investigation. Section III reports the simulation results in PSpice and those obtained by the PSpice-Simulink co-simulation, while Section IV shows analogous results obtained by the PSIM simulation and by the PSIM-Simulink co-simulation. Finally, conclusions are reported.

2. System Under Test
   The system under investigation is a three-phase Voltage Source Converter (VSC) employing six bidirectional switches and working as active rectifier feeding a dc-bus, whose voltage is \( V_{dc} \) (Fig. 1). The VSC is connected on its AC side at the Point of Common Coupling (PCC) through an LCL-filter [7], able to mitigate the harmonic distortion; moreover, the LCL-filter is connected to the three-phase voltage source by means of a transformer.

   With regard to the control part, the dc-bus voltage, \( V_{dc} \), is controlled by means of the power control. In fact, in order that the load absorbs a constant power, the voltage across the load has to remain constant (the reference voltage is \( V_{ref} = 700 \text{V} \)). An increase or a decrease of the \( V_{dc} \) implies a consequent increase or decrease of the required power. Then, the reference values for the control of the line currents will be consequently modified and the signals to the gates of the switches will be determined. A PI controller has been utilized for the voltage control as well as for the line currents control.

   The PI controller for the dc-bus voltage \( V_{dc} \) is
   \[
   G_i(s) = k_{pv} + \frac{k_i}{s} \tag{1}
   \]
   as well as for the line currents control is
   \[
   G_i(s) = k_{il} + \frac{k_i}{s} \tag{2}
   \]

   In the Table I all the parameters of the power part as well as of the control part are reported.
3. PSpice Simulation And PSpice-Simulink Co-simulation

The electric system of Fig. 1 has been implemented in PSpice environment and a transient analysis of 500ms has been considered. Fig. 2 and Fig. 3 report the waveforms of the load voltage $V_{dc}$ and of the line currents, respectively; particularly, the last ones are phase-shifted of 120° each other as expected.

Table II reports some simulation parameters of PSpice and their meanings. In order to have an efficient simulation [4], the ratio NUMRTP/NUMTTP has to be minus than 10%, but in this case it results 18.9%. Moreover the overall simulation time is very expensive (~14.5 hours) utilizing a PC Pentium 4, 3 GHz, 1GB RAM.

### Table II. Transient Analysis Parameters [3]

| NUMIT (Total number of iterations during transient simulation) | 633 506 622 |
| NUMTTP (Number of Spice internal time-points during the transient analysis) | 338 689 |
| NUMRTTP (Number of times Spice fails to converge, cuts the time-step and reattempt the circuit solution during transient analysis) | 63 905 |
| Total Job Time [s] (Seconds needed to Spice to complete overall simulation) | 52 246 |

In order to optimize the simulation performances a PSpice-Simulink co-simulation has been implemented in
Simulink environment. The interface tool SLPS allows the data exchange after defining the input and output variables, respectively. The three modulating signals are considered as input variables to SLPS, while the output voltage Vdc as well as the line currents and voltages are imposed as output variables from SLPS. The co-simulation is entirely managed in Simulink environment.

Two types of co-simulation have been considered. The former one is based on fixed-step integration methods, while the latter one is based on variable step integration methods.

Both the series of simulations have shown difficulties. In the first case, convergence problems happen after 19 ms (it needs ~15 min); even if the integration method is changed and/or tolerance values are re-set, the simulation is stopped. Nevertheless, the results obtained until the interruption agree with those expected.

In the second case, memory overflow problems happen after 40ms (it needs ~22 min); the same problems exist even considering a more powerful PC. Even in this case, the waveforms are those expected. Fig. 4 reports a zoom of the output voltage Vdc for both the last cases in the range [4÷18]ms; the waveform of Vdc for the previous simulation in PSpice environment is reported too and they are almost superimposed.

The results highlight that the simulation based only on PSpice is extremely slow, but provides expected results for the whole simulation time, whereas the co-simulation with Simulink toolbox is not adequate because convergence problems or memory overflow happen.

4. PSIM Simulation And PSIM-Simulink Co-simulation

To implement a power electronic circuit such as that of Fig. 1 in PSIM environment is more immediate than in PSpice one. This happens because PSIM software is just devoted to the analysis of power electronics, and then some typical devices are directly implemented or are easy to handle. For example, in PSIM the PI controller is constituted by one block in which the values of the gain and of the time constant are directly inserted, whereas PSpice has not an equivalent component. Moreover, PSIM has components with the role of current or voltage sensor, which are very useful to manage the control part of the circuit; in PSpice the analogue function has to be treated by means of controlled sources.

Nevertheless PSIM utilizes a fixed-step integration method (trapezoidal rule), whose time step has to be defined by the user. This condition represents a constraint for the co-simulation, because only fixed-step integration methods with the same time-step can be selected in Simulink toolbox.

In order to compare the simulation results of PSIM simulator with the previous ones, the same total time of simulation (500ms) for the transient analysis has been set, after fixing a 1µs time-step. The obtained waveform for the output voltage Vdc is globally similar to that of Fig. 2, even if they differ in the first part, as reported in Fig. 5.

It is worth highlighting the persistent fluctuations of the PSIM-waveform around the reference value 700V. Probably this fluctuation is not a true behavior of the output voltage, but it depends on the integration method (trapezoidal rule) utilized by the PSIM simulator. The authors are just investigating this issue.

Fig. 6 reports the same waveforms in the range [441÷443]ms, i.e. almost at the end of the simulation; the behavior is just alike, apart a very little offset which decreases as the simulation run.

The time taken by PSIM to complete the simulation (500ms) is 29 seconds; this is just a very little time if compared with the time employed by PSpice for the same simulation (~15 hours). Nevertheless PSIM does not allow to take other information to evaluate the simulation quality. In fact, it is not possible to take information about the total
number of iterations, the number of times it fails to converge, etc.

Summarizing, the simulation in PSIM environment is fast, but performances indicators are not available. Moreover some results have to be validated, because it seems they have affected by characteristic errors of the integration method. The choice to implement a only one integration method, with fixed step, constitutes a limit of PSIM.

As just introduced, even the PSIM simulator can interact with Simulink toolbox by means of an interface block, called SimCoupler Module. Even in this case the first step is to define the input and the output variables, respectively; the same choices of the previous case has been done. Obviously, only co-simulations with fixed-step integration methods can be run and the same time-step has been set in both PSIM and Simulink.

As in the previous cases, a 500 ms simulation with a 1µs time-step has been considered and Runge-Kutta method has been imposed in Simulink. The output voltage \( V_{dc} \) is reported in Fig. 7, in which the previous simulations are reported too. It is observed the output voltage \( V_{dc} \) is almost superimposed to that one obtained in PSpice environment, while differs by that one of PSIM. This result supports the hypothesis about the persistent fluctuation as anomaly related to the utilized integration method.

The line currents constitute a direct balanced three-phase system, as reported in Fig. 3.

Finally, the whole simulation is completed in 32 seconds. In the opinion of the authors the co-simulation PSIM-Simulink guarantees the best performances with regard to the determined variables as well as with respect to the employed time.

As shown in Fig. 8, the waveform of the phase (a) shows a phase shift around 90°, which is equal to the converter's behavior as a resistor, as expected. Moreover both the waveforms are almost sinusoidal, i.e. the LCL filter is able to contain the harmonic distortion produced by the converter operation. Both these requirements are strictly needed in order to design a well-behaving boost active rectifier.

Obviously, analogous waveforms are obtained for the other two phases (b) and (c).

Fig. 8 reports the waveforms of both the line current and voltage for the phase (a). The same behavior is obtained for the PSpice simulation. It is worth noting two aspects: the former one regards the phase shift between the waveform, the latter one regards the harmonic distortion.

The phase shift is almost equal to zero, i.e. the converter behaves as a resistor, as expected.

Moreover both the waveforms are almost sinusoidal, i.e. the LCL filter is able to contain the harmonic distortion produced by the converter operation. Both these requirements are strictly needed in order to design a well-behaving boost active rectifier.

Fig. 7 reports the waveforms of both the line current and voltage for the phase (a). The same behavior is obtained for the PSpice simulation. It is worth noting two aspects: the former one regards the phase shift between the waveform, the latter one regards the harmonic distortion.

The phase shift is almost equal to zero, i.e. the converter behaves as a resistor, as expected.

Moreover both the waveforms are almost sinusoidal, i.e. the LCL filter is able to contain the harmonic distortion produced by the converter operation. Both these requirements are strictly needed in order to design a well-behaving boost active rectifier.

Obviously, analogous waveforms are obtained for the other two phases (b) and (c).

4. Conclusions

This paper presents the simulation results for the transient analysis of switching circuits. As very different time constants are present in the circuit, convergence issues happen. The first set of simulations has been run under a general-purpose software (PSpice); after this, a co-simulation PSpice-Simulink has been run to better manage the previous issue about the time constants. Both of them have shown great problems, related to the simulation time or to the convergence or to the memory overflow, if the system under test is rather complex. The second set of simulations has been run under devoted software (PSIM); even in this case a co-simulation PSIM-Simulink has been considered. Discrete results have been obtained in PSIM environment, while the best performances are observed with the co-simulation PSIM-Simulink. The research is not closed and other complex systems will be analyzed following the same methodology.

Fig. 7. Zoom of the voltages \( V_{dc} \) deriving by PSpice, PSIM and PSIM-Simulink simulations in the range \([0\ldots125]\)ms

Fig. 8. Line current and voltage of the phase (a)

References
Reduced Bifurcation Diagram for Harmonic Balance Method using Invariant
Masakazu Yagi† and Takashi Hisakado‡
†‡Department of Electrical Engineering, Kyoto University
Kyotodaiagakusatura Nisikyo, Kyoto, 615-8510, Japan
Email: †yagi@circuiit.kuee.ykoto-u.ac.jp, ‡hisakado@kuee.kyoto-u.ac.jp

Abstract—This paper presents an efficient method to obtain a bifurcation diagram for harmonic balance (HB) method using Gröbner base based on symmetry. Gröbner base is a powerful tool to obtain the bifurcation diagram. However, its computation is very time-consuming when the equation has equivalent different solutions based on the symmetry. We show that using invariants enables us to transpose the equivalent different solutions to a unique solution. As a result, the relation between the amplitudes of each frequency component does not depend on the source using the invariants based on the symmetry of the HB equation without source.

1. Introduction
The harmonic balance (HB) method is known as the method to analyze periodic oscillations for nonlinear circuit systems. Using this method, the circuit equations are transposed to simultaneous algebraic equations which are called HB equations due to the approximation by a truncated Fourier series of variables. We have been applied the HB method to the analysis of bifurcation phenomena because the method provide the analytical results simpler than the time-domain techniques [1, 2, 3].

The harmonic balance (HB) method is known as the method to analyze periodic oscillations for nonlinear circuit systems. Using this method, the circuit equations are transposed to simultaneous algebraic equations which are called HB equations due to the approximation by a truncated Fourier series of variables. We have been applied the HB method to the analysis of bifurcation phenomena because the method provides the analytical results simpler than the time-domain techniques [1, 2, 3].

The HB equation without source has the symmetry based on an arbitrary phase shift. The invariants with respect to this symmetry correspond to the amplitudes of the each frequency component. This paper presents also an efficient method to obtain the relation between the amplitudes using the invariants. The relation does not depend on the source since it is obtained by the HB equation without source.

2. Harmonic Balance (HB) Equation
We consider a circuit equation of n dimensions
\[
\frac{du(t)}{dt} = h(u, \lambda) + e(t),
\]
where \( h(u, \lambda) = (h_1, \ldots, h_n)^T \in \mathbb{R}^n, u(t) = (u_1, \ldots, u_n)^T \in \mathbb{R}^n, \lambda = (\lambda_1, \ldots, \lambda_n) \in C^n, \) \( e(t) = (e_1, \ldots, e_n)^T \in \mathbb{R}^n, \) and \( u \) is a state variable vector, \( e(t) \) is a source vector of period \( 2\pi/(m \in \mathbb{Z}_{>0}), \lambda \) is a set of circuit parameters, \( \alpha \) is an integer.

The HB method is known as the method to analyze periodic oscillations for nonlinear circuit systems. Using this method, the circuit equations are transposed to simultaneous algebraic equations which are called HB equations due to the approximation by a truncated Fourier series of variables. We have been applied the HB method to the analysis of bifurcation phenomena because the method provides the analytical results simpler than the time-domain techniques [1, 2, 3].

The HB equation without source has the symmetry based on an arbitrary phase shift. The invariants with respect to this symmetry correspond to the amplitudes of the each frequency component. This paper presents also an efficient method to obtain the relation between the amplitudes using the invariants. The relation does not depend on the source since it is obtained by the HB equation without source.
satisfy the relation and have equivalent different solutions. The number of equivalent asymmetric solutions for $\Gamma_{2n/m}$ is $m$.

If the nonlinear characteristics of the circuit equation has odd symmetry and the source satisfies $e(\tau + \pi) = -e(\tau)$, the system has another symmetry described by

$$f(\Theta_{\gamma} \psi, \lambda) = \Theta_{\gamma} f(\psi, \lambda), \quad i = 1, 2,$$

where

$$\Theta_{\gamma} = \begin{bmatrix} I_{2n} & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & \cdots & \cdots & \cdots & \cdots & I_{2n} \end{bmatrix},$$

$k = 0, \ldots, p$ and $I_{2n}$ is an identity matrix of $2n \times 2n$. We call the symmetry of $\pi$ phase shift and sign inversion $\Gamma_{\pi}$. The symmetric solutions with respect to $\Gamma_{\pi}$ satisfy $\psi = \Theta_{\pi} \psi$ for all $i$, whereas asymmetric solutions do not satisfy the relation. The number of equivalent asymmetric solutions for $\Gamma_{\pi}$ is 2. Thus, the oscillation which is asymmetric for $\Gamma_{2n/m}$ and $\Gamma_{\pi}$ has 2$m$ equivalent different solutions.

### 3.2. Bifurcation Diagram using Invariant

Since the HB equation is a simultaneous algebraic equation, we can obtain the bifurcation diagram described by a unique polynomial equation using Gröbner base. For example, the Gröbner base with respect to lexicographic order $\psi_{01} > \cdots > \psi_{3p,n}$ gives a bifurcation diagram $(\lambda - \psi_{3pe})$. However, it is known that the circuit is generally very time-consuming. Especially, if the equation has equivalent different solutions, the degree of the polynomial becomes high and the difficulty of the computation increases.

We propose to decrease the degree of the bifurcation diagram using invariant which enables to transpose the different equivalent solutions to a unique solution.

An invariant $g(\psi)$ with respect to a symmetry $\Gamma$ is defined by

$$g(\psi) = g(\Theta(\gamma) \psi),$$

for all $\gamma \in \Gamma$. We consider the set of the invariants as follows:

$$E(\rho)^I := \{ f \in E(\rho) \mid \forall \gamma \in \Gamma, f = \Theta(\gamma) f \}. \quad (7)$$

Because this set is closed under addition and multiplication, it is a sub-ring of $E(\rho)^I$. $E(\rho)^I$ is called the invariant sub-ring of $\rho$. It has been known that the arbitrary invariant $g \in E(\rho)^I$ can be rewritten by finite fundamental invariants $g_1, \ldots, g_l$ [8]. If we define variables $A_{ij}$, $\gamma_i \equiv g_i(\psi), i = 1, \ldots, r.$

The equivalent asymmetric solutions are represented by a unique solution $\lambda_i$.

We obtain the bifurcation diagram by the ideal $(f(\psi, \lambda))$ using Gröbner base of lexicographic order. On the other hand, the reduced bifurcation diagram by the invariant is obtained by the ideal $(f(\psi, \lambda)) + \sum (g_i - A_i)$. We propose algorithm to obtain the reduced bifurcation diagram described as follows:

**S1.** We give invariants $A_{01}, \ldots, A_{rs}$, which are the targets for bifurcation diagram, where $a_1, \ldots, a_s \leq r$.

**S2.** We obtain the Gröbner base of lexicographic order $\psi > A_{01} > \cdots > A_{as}$ by the ideal $(f(\psi, \lambda)) + (g_{a1} - A_{a1}, \ldots, g_{as} - A_{as})$.

**S3.** We can obtain the reduced bifurcation diagram by the the eliminated ideal which contains only $A_{01}, \ldots, A_{as}$ in the Gröbner base.

The degree of the bifurcation diagram is decreased by this method and the computation time of the Gröbner base is also decreased.

### 4. Example

#### 4.1. Circuit Equation

We apply the proposed method to an RLC resonance circuit shown in Figure 1. The characteristics of the nonlinear inductor is approximated by cubic polynomial, i.e., $e(\phi) = \psi \phi^2$, where $\phi$ is a magnetic flux and $i$ is current. The scaled circuit equation is written by

$$\frac{du(\tau)}{d\tau} = -k(c_1 u_1(t) + c_3 u_3(t)) + E \cos m\tau \quad (9)$$

where $c_1 = 1 - c_3 = \frac{m}{\sqrt{L_0}}, c_3 = \frac{m^2}{\sqrt{L_0 C}}, k = \frac{mE}{\sqrt{C}}, E = \frac{C E_0}{\sqrt{L_0}}$, where $L_0$ is the value of the normalized current, and $v_i$ is the capacitance voltage. Using the HB method with 3 frequency components, the $u_1(\tau)$ and $u_2(\tau)$ are respectively approximated by

$$u_1(\tau) = \psi_{31}, \psi_{32} \cos 2\pi t + \psi_{33} \cos 3\pi t,$$

$u_2(\tau) = \psi_{21} \sin 2\pi t + \psi_{22} \sin 2\pi t + \psi_{23} \cos 2\pi t + \psi_{24} \cos 3\pi t, \quad (10)$

where $u_2$ is obtained by integral of $u_1^*$ for simplification, i.e., $\psi_{21,1}, \psi_{21,2}, \psi_{22,1}, \psi_{22,2}, \psi_{23,1}, \psi_{23,2}, \psi_{24}$ are represented by $\psi \equiv (\psi_{01}, \psi_{02}, \psi_{03}, \psi_{04})$, and the DC component $\psi_{00}$ is zero. We consider the case of $m = 3$, i.e., the target oscillation is the 1/3 sub-harmonic oscillation. The HB equation is written by $f(\psi, \lambda) = (f_1, f_{11}, f_{21}, f_{22}, f_{23}, f_{31}) = 0$, where $A \equiv (c_3, k)$.

#### 4.2. Symmetry of the System

In order to obtain the fundamental invariants, let us consider the symmetry. The HB equation $f(\psi, \lambda) = 0$ has the symmetries $\Gamma_{2\pi/3}$ and $\Gamma_{\pi}$. The corresponding linear representation $\Theta_{2\pi/3}$ and $\Theta_{\pi}$ are

$$\Theta_{2\pi/3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{2} & \frac{\sqrt{3}}{2} \\ -\frac{1}{2} & \frac{\sqrt{3}}{2} & 1 \end{bmatrix}, \quad \Theta_{\pi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (12).$$

The number of equivalent asymmetric solutions for $\Gamma_{2\pi/3}$ is 3, and the number of equivalent asymmetric solutions for $\Gamma_{2\pi/3}$ and $\Gamma_{\pi}$ is 6.

#### 4.3. Invariants of the System

Using the $\Theta_{2\pi/3}$ and $\Theta_{\pi}$, the fundamental invariants with respect to the symmetries $\Gamma_{2\pi/3}$ and $\Gamma_{\pi}$ are

$$g(\psi) = \psi_{01}^2 + \psi_{02}^2, \quad (13)$$
Solution 1
Solution 2
Solution 3

Figure 2: The parameter E versus the amplitude of the fundamental oscillation \( \sqrt{A_3} \). (A = (c_3, k) = (1,1/150)). “Solution 1” represents asymmetric solutions, “Solution 2” represents solutions based on symmetry \( \Gamma_{12} \), and “Solution 3” represents solutions based on symmetries \( \Gamma_{23} \) and \( \Gamma_{24} \).

\[
g_2(\psi) = \psi_{11} (3\psi_{21}^2 - \psi_{11}^2),
\]
(14)

\[
g_3(\psi) = \psi_{11} (\psi_{21}^2 - 3\psi_{11}^2),
\]
(15)

\[
g_4(\psi) = \psi_{22}^2 + \psi_{12}^2,
\]
(16)

\[
g_5(\psi) = \psi_{22} (3\psi_{22}^2 - \psi_{11}^2)(\psi_{22}^2 - 3\psi_{11}^2),
\]
(17)

\[
g_6(\psi) = \psi_{22}^2 + 21\psi_{22}^2\psi_{12}^2 - 9\psi_{11}^2\psi_{22}^2 + 3\psi_{12}^2,
\]
(18)

\[
g_7(\psi) = \psi_{22}^6 - 3\psi_{22}^2\psi_{12}^2 + 21\psi_{12}^2\psi_{22}^2 + \psi_{12}^2.
\]
(19)

where \( g_1(\psi), g_2(\psi), g_3(\psi) \) enable to transpose the equivalent 3 asymmetric solutions for \( \Gamma_{23} \) to a unique solution and \( g_4(\psi), \ldots, g_7(\psi) \) enable to transpose the equivalent 6 asymmetric solutions for \( \Gamma_{23} \) to a unique solution. The \( g_i(\psi) \) and \( g_j(\psi) \) correspond to the amplitudes of 1/3 and 2/3 subharmonic frequency components. Thus, all arbitrary invariants can be rewritten by \( g_i(\psi), i = 1, 2, \ldots, 7 \).

Let us consider the bifurcation diagram of invariants which correspond to the amplitudes of each frequency component. We rewrite invariants as follows:

\[
A_1 \equiv g_1(\psi) = \psi_{11}^2 + \psi_{12}^2,
\]
(20)

\[
A_2 \equiv g_2(\psi) = \psi_{22}^2 + \psi_{12}^2,
\]
(21)

\[
A_3 \equiv g_3(\psi) = \psi_{22}^2 + \psi_{12}^2.
\]
(22)

The \( g_3(\psi) \) is an invariant since all polynomials of the fundamental oscillation components \( \psi_{11}, \psi_{12} \) and \( \psi_{22} \) are invariants. \( \sqrt{A_1}, \sqrt{A_2} \) and \( \sqrt{A_3} \) are the amplitudes of each frequency component.

### 4.4. Computation Efficiency

In order to confirm the efficiency of the proposed method, we compare the computation time between two methods, i.e., the method with and without the invariants. We use computer algebra system Risa/Asir [9].

In the method without the invariants, we consider to obtain the bifurcation diagram by ideal \( (f(\psi, \lambda)) \) using Gröbner base of lexicographic order \( \psi_{12} > \psi_{22} > \psi_{11} > \psi_{3} > \psi_{12} \). i.e., we obtain the bifurcation diagram \( (\theta - \psi_{12}) \) with parameters \( \lambda = (c_3, k) \). However, its computation requires more than 4GB memory and we cannot obtain the bifurcation diagram.

However, using the proposed method with the invariants, we can obtain the reduced bifurcation diagram \( (E - A_3) \) by ideal \( f(\psi, \lambda) + (g_1 - A_1, g_2 - A_2, g_3 - A_3) \) using Gröbner base of lexicographic order \( \psi_{12} > \psi_{22} > \psi_{11} > \psi_{3} > \psi_{12} > A_2 > A_1 > A_3 \). The reduced bifurcation diagram can be decomposed to 3 factors — corresponding to asymmetric solutions (Solution 1), solutions based on symmetry \( \Gamma_{12} \) (Solution 2) and solutions based on symmetries \( \Gamma_{23} \) and \( \Gamma_{24} \) (Solution 3) as shown in Figure 2 where the set of parameters \( \lambda = (c_3, k) = (1,1/150) \).

We consider a simpler problem for the comparison. We set the 2/3 subharmonic frequency components \( \psi_{22} = \psi_{12} = 0 \), i.e., we consider the HB method with 2 frequency components, 1/3 subharmonic and fundamental harmonic components. In this case, the number of equivalent asymmetric solutions for \( \Gamma_{23} \) is 3. Each bifurcation diagram \( (E - \psi_{12}) \) and \( (E - A_1) \) can be obtained using lexicographic order \( \psi_{12} > \psi_{22} > \psi_{11} > \psi_{3} > A_3 > A_1 \). The results are shown in Table 1 and Table 2.

In Table 1, the number of the asymmetric solutions except symmetric solution \( \psi_{11} = 0 \) is 18 in the case without the invariants, whereas the number of the asymmetric solutions except symmetric solution \( A_1 = 0 \) is 6 in the case with the invariants. We can confirm that the invariant \( A_1 \) transposes the equivalent 3 asymmetric solutions to a unique solution.

The number of terms when using the method with the invariants is considerably smaller than the number of terms on the method without the invariants. The Table 2 compare the computation time. We can confirm the efficiency of the proposed method.

Thus, the bifurcation diagram is reduced by the invariants and its computation time is decreased dramatically.

### 5. Relation between Amplitudes using Invariant

Let us consider the following HB equation without source from the circuit equation (1)

\[
\dot{\psi}(\lambda) = (\dot{f}_0, \dot{f}_1, \dot{f}_2, \ldots, \dot{f}_p, \dot{e}_p) = 0 \in \mathbb{R}^N, (23)
\]

\[
\dot{e}_p = \dot{f}_p + \dot{e}_p, \quad (\dot{e}_0, \dot{e}_1, \dot{e}_2, \ldots, \dot{e}_p) \in \mathbb{R}^N
\]

The HB equation (3) is written by \( \dot{f}(\psi, \lambda) = \dot{\psi}(\lambda) + \dot{\psi}, \dot{\psi}_1 = 0 \). In many cases, the Eq.(3) contains the equations which do not depend on the source \( \dot{\psi} \) because many elements of the \( \dot{\psi} \) are zeroes.

For example, we consider the case in the previous section. The cases of \( m = 1, 2, \) and \( 3 \) correspond to fundamental oscillation, 1/2 subharmonic oscillation and 1/3 subharmonic oscillation for the HB, respectively. The HB equations \( \dot{f}(\psi, \lambda) \) of \( m = 1, 2, \) and \( 3 \) are represented as

\[
\dot{f}(\psi, \lambda) = (\dot{f}_1 + E, \dot{f}_2 + E, \dot{f}_3 + E, \dot{f}_4 + E)^T = 0, (24)
\]

\[
\dot{f}(\psi, \lambda) = (\dot{f}_1, \dot{f}_2 + E, \dot{f}_3 + E, \dot{f}_4 + E)^T = 0, (25)
\]

\[
\dot{f}(\psi, \lambda) = (\dot{f}_1, \dot{f}_2, \dot{f}_3 + E, \dot{f}_4 + E)^T = 0, (26)
\]

Thus, the source \( E \) of frequency \( m \) affects only the HB equation of the corresponding frequency. So the equations

<table>
<thead>
<tr>
<th>Method</th>
<th>Max degree</th>
<th>Number of terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>without Invariant</td>
<td>19</td>
<td>64</td>
</tr>
<tr>
<td>with Invariant</td>
<td>7</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Computation time [s]</th>
</tr>
</thead>
<tbody>
<tr>
<td>without Invariant</td>
<td>6.14306</td>
</tr>
<tr>
<td>with Invariant</td>
<td>0.6509</td>
</tr>
</tbody>
</table>

| Table 2: Computation time of the bifurcation diagram. |
|--------|----------------------|
| \( \lambda = (1, k) \) | \( \lambda = (1, 1/150) \) |
| without Invariant | 6.14306 | 25000.9 |
| with Invariant | 0.6509 | 0.7209 |
which do not contain $E$ have a symmetry of the HB equation without source $(23)$.

The Eq.(23) has the symmetry $\Gamma_\xi$ as follows:

$$\hat{f}(\Theta_\xi \phi, \lambda) = \Theta_\xi \hat{f}(\phi, \lambda),$$  

(27)

where

$$\Theta_\xi = \begin{bmatrix} \theta_{\xi,0} & 0 \\ 0 & \Theta_{\xi,m} \end{bmatrix}_m^{n \times 2n},$$

and $k = 0, \ldots, p$. This symmetry $\Gamma_\xi$ is based on the arbitrary phase $\xi$ shift. The invariants $A_{p1} = g_{p1}, \ldots, A_{p5} = g_{p5}$ with respect to $\Gamma_\xi$ correspond to the amplitudes.

Using the invariants $g_{p1} - A_{p1} = 0, \ldots, g_{p5} - A_{p5} = 0$ and equations which do not depend on the source $\hat{e}$ in Eq.(3), we can obtain the relation between the amplitudes of the each frequency component because the $A_{p1}, \ldots, A_{p5}$ do not depend on phase. The relation does not depend on source, since the invariants and HB equation without source $(23)$ do not depend on the source.

In the case of the example, the invariants $A_1 = \phi_{11}^2 + \phi_{12}^2, A_2 = \phi_{21}^2 + \phi_{22}^2, A_3 = \phi_{31}^2 + \phi_{32}^2$ as shown in Figure 4. In case of $m = 1$, the relation between amplitudes $A_1, A_2, A_3$ are obtained by $f_{\xi,1}, f_{\xi,2}, f_{\xi,3}$ as shown in Figure 4. In case of $m = 2$, the relation between amplitudes $A_1, A_2, A_3$ are obtained by $f_{\xi,1}^2, f_{\xi,2}^2, f_{\xi,3}^2$ as shown in Figure 4. In case of $m = 3$, the relation between amplitudes $A_1, A_2, A_3$ are obtained by $f_{\xi,1}^3, f_{\xi,2}^3, f_{\xi,3}^3$ as shown in Figure 5. The solutions are on the dotted lines in Figure 3 and Figure 5, the $\phi_{\xi,1}$ and $\phi_{\xi,2}$ are obtained as complex numbers since $A_3$ is a negative number.

Thus, we can obtain the relation between amplitudes of the each frequency component by the HB equation without source using invariants. It is noted that the relation does not depend on the source.

6. Conclusion

This paper proposed an efficient method to obtain the reduced bifurcation diagram using the invariants based on the symmetry. In this method, the degree of the bifurcation diagram is decreased because the equivalent different solutions are transposed to a unique solution by the invariants. We confirmed that the proposed method considerably decrease the computation time. Further, based on symmetry of the HB equation without source, we presented a method to obtain the relation between the amplitudes of the each frequency which does not depend on the source components using the invariants.

Acknowledgment

One of the authors expresses his sincere thanks to the professor emeritus of Kyoto University, Kohshi Okumura who has led his way. This work was supported by the 21st Century COE Program 14213201.

References


On Optimization Algorithm for Attaining The Maximum DC Gain of CMOS Amplifiers

M. Oda†, Y. Yamagami‡, Y. Nishio†, J. Kawata‡, A. Ushida‡

† Dept. of Electrical and Electronic Engineering, Tokushima University, Tokushima 770-8506, Japan
‡ Dept. of Mechanical and Electronic Engineering, Tokushima Bunri University, Kagawa 769-2193, Japan

Abstract
CMOS operational amplifiers are most important building blocks of analog circuits. In this paper, we propose a Spice-oriented design algorithm of CMOS amplifiers for attaining the maximum DC gain. Our optimization algorithm is based on well-known steepest descent method, where the gradient direction is decided by the analysis of sensitivity circuits for the designing parameters such as MOS sizes and bias voltages. In order to develop the user-friendly designing tool, we have firstly tried to replace MOSFETs by the corresponding sensitivity modules. After then, our steepest descent algorithm is realized by equivalent nonlinear RC circuits with ABMs (analog behavior models), so that the solution can be found as an equilibrium point with transient analysis of Spice. We found from some numerical experiments that the optimum points can be stably obtained by the transient analysis.

1. Introduction
CMOS operational amplifiers are widely used as the building blocks in analog circuits [1-4]. For designing the amplifiers, there are several design criteria such as the maximum gain, the minimum power consumption and so on. For example of attaining the maximum gain of an amplifier, the gain largely depends on the bias voltages and MOS sizes, so that they are chosen as the design parameters. Traditionally, the optimum parameters were found by trial and error process with Spice. It was really time-consuming when the design parameters are increased [5]. On the other hand, the steepest descent method is widely used for this kind of problems [6]. However, it is very efficient only when the objective functions are given in the analytical form.

Hence, in this paper, we propose a practical steepest descent algorithm based on Spice combining the sensitivity analysis. Firstly, we need to define the objective function as follow:

\[ \Phi(x, p), \ x \in \mathbb{R}^n, \ p \in \mathbb{R}^k \] (1)

where
- \( x \): circuit variables such as voltages and currents.
- \( p \): optimization parameters such as bias voltages, resistor’s values and the dimensions of MOSFETs W[width in \( \mu m \)] and L[length \( \mu m \)] and so on [6,7]

Now, let us discuss the design algorithm of amplifiers for attaining the maximum DC gain based on Spice-oriented steepest descent method [5,7]. Namely, the gradient direction is decided by the solutions of sensitivity circuits [8], and the steepest descent algorithm can be realized by the equivalent RC circuits combining with nonlinear controlled current sources. In order to develop user-friendly simulators, we tried to transforms MOSFETs into the corresponding sensitivity modules. The optimization algorithm to attain the maximum gain is given by

\[ \frac{dp_i}{ds} = -\frac{dS_i(p)}{dp_i}, \quad i = 1, 2, \ldots, k, \] (2)

where \( S(p) \) is the sensitivity gain and \( p_i \) are optimization parameters. Unfortunately, it is impossible to evaluate \( \frac{dS(p)}{dp_i} \) from the sensitivity analysis, so that we need to introduce the numerical differentiation as follows:

\[ \frac{dp_i}{ds} \approx \frac{S_i(p + \Delta p) - S_i(p)}{\Delta p_i}, \quad i = 1, 2, \ldots, k, \]

\[ \Delta p = (0, 0, \ldots, \Delta p_i, \ldots, 0)^T \] (3)

with sufficiently small \( \Delta p \). Then, replacing a variable “s” by time “t”, our descent algorithm can be realized by the equivalent nonlinear RC circuits with the solutions of the sensitivity circuit and ABMs of Spice[9]. In this way, we can find out the optimum parameters at the equilibrium point from the transient analysis.

We show the sensitivity circuit using the tableau approach, and the above optimization algorithm in section 2. The sensitivity modules of MOSFETs are shown in section 3. They are very useful to develop user-friendly simulators. We show interesting examples of inverter amplifier to attain the maximum gain and CMOS amplifier in section 4.

2. Sensitivity analysis and Spice-oriented optimization algorithm

Sensitivity analysis
The steepest descent method is the most basic optimization approach, where the gradient direction is decided by solutions of the sensitivity circuit. Let us derive the sensitivity circuit via tableau approach [8]. The tableau equation is given by:

\[
\begin{bmatrix}
K_i & K_v & 0 \\
0 & 1 & -A^T \\
A & 0 & 0
\end{bmatrix}
\begin{bmatrix}
i \\
v \\
v_n
\end{bmatrix}
- \begin{bmatrix}
g(v, i) \\
E \\
AJ
\end{bmatrix}
= \begin{bmatrix}
0 \\
0 \\
0
\end{bmatrix}
\]

(4)

The first row is coming from the Ohm’s law, second and third are Kirchhoff’s voltage and current lows, respectively. \( A \) is the incidence matrix and \( K_i, K_v \) are \((b \times b)\) square matrix composed of “1” and “0” for \( b \) elements. \( E \) and \( J \)
show the voltage and current sources. Now, we define the sensitivity as follow:

**Sensitivity:** Let us define the sensitivity by

$$S_{k,i} = \lim_{\Delta p_i \to 0} \frac{\Delta x_k}{\Delta p_i}$$  \hspace{1cm} (5)

Now, let us calculate the variational equation from (3). For parameter variations of the nonlinear resistors, we have

$$g(v_0 + \Delta v, i_0 + \Delta i) + \frac{\partial g(v,i)}{\partial p} \Delta p = 0$$

$$g(v_0, i_0) = \frac{\partial g(v,i)}{\partial v} \Delta v + \frac{\partial g(v,i)}{\partial i} \Delta i$$

Thus, we have the following circuit equation for calculating the sensitivities:

$$\begin{bmatrix} K_i & K_v & 0 \\ 0 & 1 & -A^T \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{v,p_i} \\ S_{i,p_i} \\ S_{v,i} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(v,i)}{\partial v} & 0 & 0 \\ 0 & \frac{\partial g(v,i)}{\partial i} & 0 \\ 0 & 0 & \frac{\partial g(v,i)}{\partial p} \end{bmatrix} \begin{bmatrix} \delta(v_0, v_{0,k}) \\ \delta(i_0, i_{0,k}) \\ \delta(p_0, p_{0,k}) \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} K_i & K_v & 0 \\ 0 & 1 & -A^T \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{v,p_i} \\ S_{i,p_i} \\ S_{v,i} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(v,i)}{\partial v} & 0 & 0 \\ 0 & \frac{\partial g(v,i)}{\partial i} & 0 \\ 0 & 0 & \frac{\partial g(v,i)}{\partial p} \end{bmatrix} \begin{bmatrix} \delta(v_0, v_{0,k}) \\ \delta(i_0, i_{0,k}) \\ \delta(p_0, p_{0,k}) \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} K_i & K_v & 0 \\ 0 & 1 & -A^T \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{v,p_i} \\ S_{i,p_i} \\ S_{v,i} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(v,i)}{\partial v} & 0 & 0 \\ 0 & \frac{\partial g(v,i)}{\partial i} & 0 \\ 0 & 0 & \frac{\partial g(v,i)}{\partial p} \end{bmatrix} \begin{bmatrix} \delta(v_0, v_{0,k}) \\ \delta(i_0, i_{0,k}) \\ \delta(p_0, p_{0,k}) \end{bmatrix}$$

$$\Rightarrow 1 \begin{bmatrix} K_i & K_v & 0 \\ 0 & 1 & -A^T \\ A & 0 & 0 \end{bmatrix} \begin{bmatrix} S_{v,p_i} \\ S_{i,p_i} \\ S_{v,i} \end{bmatrix} = \begin{bmatrix} \frac{\partial g(v,i)}{\partial v} & 0 & 0 \\ 0 & \frac{\partial g(v,i)}{\partial i} & 0 \\ 0 & 0 & \frac{\partial g(v,i)}{\partial p} \end{bmatrix} \begin{bmatrix} \delta(v_0, v_{0,k}) \\ \delta(i_0, i_{0,k}) \\ \delta(p_0, p_{0,k}) \end{bmatrix}$$

$$\Rightarrow i = c_1 v + c_3 v^3$$

is chosen as an optimization parameter. Then, the sensitivity circuit is given by Fig.2(b).

Fig.1. Sensitivity elements for resistors.

Fig.2. Sensitivity circuit.

**Example:** Now, consider a simple example for deriving the sensitivity circuit. We assume that $c_3$ in nonlinear resistor

$$i = c_1 v + c_3 v^3$$

is chosen as an optimization parameter. Then, the sensitivity circuit is given by Fig.2(b).

**Variational method:** We perturb $c_3 \rightarrow c_3 + \Delta c_3$, then, we have the following circuit equation and the variational equation:

$$c_1(v_0 + \Delta v) + (c_3 + \Delta c_3)(v_0 + \Delta v)^3 + \frac{(v_0 + \Delta v) - E}{R_1} = 0$$

$$\Rightarrow c_1 \Delta v + c_3 \Delta v^3 + 3c_3 v^2 \Delta v + \Delta v = 0$$

Therefore, the sensitivity of the resistor voltage is given by

$$S_v = \lim_{\Delta c_3 \to 0} \frac{\Delta v}{\Delta c_3} = -\frac{v_0^3}{G(v_0) + \frac{1}{R_1}}$$

where $G(v_0) = c_1 + 3c_3 v_0^2$.

**Sensitivity analysis:** From the sensitivity circuit Fig.2(b) with $\partial i/\partial c_3|_{v_0} = v_0^3$, we have the voltage drop at the resistor $G(v_0)$ is given

$$S_v = -\frac{v_0^3}{G(v_0) + \frac{1}{R_1}}$$

which is exactly equal to the solution from the variational method.

**Spice-oriented optimization technique**

Now, we consider optimization problems of DC circuits, and define the objective function as follows:

$$\Phi(v, i, p) = \hat{\phi}(v_1, v_2, \ldots, v_n, i_1, i_2, \ldots, i_m, p_1, p_2, \ldots, p_k)$$

(8)

where $v, i$ are circuit variables, and $p$ is the optimization parameters. Applying the steepest descent method to (7), we have

$$\frac{\partial \Phi}{\partial v} = -\hat{\phi}(S_{v_1,p_1}, \ldots, S_{v_n,p_1}, S_{i_1,p_2}, \ldots, S_{i_m,p_1}, 1, 0, \ldots, 0)$$

$$\frac{\partial \Phi}{\partial i} = -\hat{\phi}(S_{v_1,p_2}, \ldots, S_{v_n,p_2}, S_{i_1,p_2}, \ldots, S_{i_m,p_2}, 0, 1, \ldots, 0)$$

$$\frac{\partial \Phi}{\partial p} = -\hat{\phi}(S_{v_1,p_k}, \ldots, S_{v_n,p_k}, S_{i_1,p_k}, \ldots, S_{i_m,p_k}, 0, 0, \ldots, 1)$$

(9)

where $\{S_{i_1,p_1}, \ldots, S_{i_j,p_j}, i = 1, 2, \ldots, n, j = 1, 2, \ldots, k\}$ are solutions from the sensitivity analysis. Note that we need $k$ times sensitivity analysis to execute the steepest descent method. Changing variable “$s$” with time “$t$”, the equation (8) is realized by a coupled RC circuits using controlled-current sources as shown in Fig.3, where the value capacitance $C_s$ is related to the convergence speed, and $R_0$s are sufficient large dummy resistances. On the other hand,
the optimization attaining the maximum gain is realized by the use of numerical differentiations as given in (2). The relation (2) is also realized by the circuit shown in Fig. 4 with small $\Delta p_i$, where $S(p + \delta(\Delta p_i))$ and $S(p)$ are so-

\[
S(p + \delta(\Delta p_i)) = S(p) + \delta S(p) \Delta p_i \quad (i = 1, 2, \ldots, k)
\]

Fig. 4. Circuit configuration of attaining the maximum gain.

3. Sensitivity modules of MOSFETs

Now, let us derive the sensitivity modules of MOSFETs in DC. Once these modules are stored in our computer as the packages, our optimization algorithms can be easily carried out with Spice. We apply piecewise continuous Shichman-Hodges model [1-4] for nMOS as follows:

1. **Linear region** ($v_{GS} > v_T, 0 < v_D < v_{GS} - v_T$)

\[
i_D = \frac{k_n W}{L} (v_{GS} - v_T - \frac{v_D}{2}) v_D (1 + \lambda v_D) \quad (10)
\]

2. **Saturation region** ($v_{GS} > v_T, v_D \geq v_{GS} - v_T$)

\[
i_D = \frac{k_n W}{2L} (v_{GS} - v_T)^2 (1 + \lambda v_D) \quad (11)
\]

where the threshold voltage is given by

\[
v_T = v_{TH} + \gamma \left( \sqrt{\phi - v_{BS}} - \sqrt{\phi} \right) \quad (12)
\]

Thus, we have the following sensitivity elements at the corresponding operating points:

1. **Linear region**

\[
G_{GS} \equiv \frac{\partial i_D}{\partial v_{GS}} = \frac{k_n W}{L} v_D (1 + \lambda v_D) \quad (13)
\]

\[
G_{DS} \equiv \frac{\partial i_D}{\partial v_D} = \frac{k_n W}{L} \times \left[ -\frac{3}{2} v_D (1 + 2\lambda) (v_{GS} - v_T) \right] \quad (14)
\]

\[
G_{BS} \equiv \frac{\partial i_D}{\partial v_{BS}} = \gamma k_n W v_D (1 + \lambda v_D) \quad (15)
\]

2. **Saturation region**

\[
G_{GS} \equiv \frac{\partial i_D}{\partial v_{GS}} = \frac{k_n W}{L} (v_{GS} - v_T) (1 + \lambda v_D) \quad (16)
\]

\[
G_{DS} \equiv \frac{\partial i_D}{\partial v_D} = \frac{k_n W}{2L} (v_{GS} - v_T) \lambda v_D \quad (17)
\]

\[
G_{BS} \equiv \frac{\partial i_D}{\partial v_{BS}} = \gamma k_n W v_D (1 + \lambda v_D) \quad (18)
\]

For the optimization parameter ($W, L$), we have the following current sensitivities:

1. **Linear region**

\[
S_{in} = \frac{k_n W}{L} \left( v_{GS} - v_T \right) v_D (1 + \lambda v_D), \quad \text{for } L \quad (19)
\]

\[
S_{in} = \frac{k_n W}{2L} \left( v_{GS} - v_T \right) v_D (1 + \lambda v_D), \quad \text{for } W, \quad (20)
\]

2. **Saturation region**

\[
S_{in} = \frac{k_n W}{L} \left( v_{GS} - v_T \right)^2 (1 + \lambda v_D), \quad \text{for } L \quad (21)
\]

\[
S_{in} = \frac{k_n W}{2L} \left( v_{GS} - v_T \right)^2 (1 + \lambda v_D), \quad \text{for } W \quad (22)
\]

where “0” means voltage at the operating point. These modules are realized with ABMs of Spice, whose nMOS module is shown by Fig. 5(b). We can also realize pMOS sensitivity modules in the same way.

4. Illustrative examples.

**CMOS inverter amplifier.**

Most of CMOS operational amplifiers contain some kinds of inverters Fig 6(a) as the building blocks, whose transfer characteristic is given by Fig. 6(b), where MOS parameters are given by Table 1.

![CMOS Inverter](image)

**Table.1 MOS parameter[9] (MKS system).**

<table>
<thead>
<tr>
<th>Mark</th>
<th>pMOS</th>
<th>nMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>$1.2 \times 10^{-6}$</td>
<td>$1.2 \times 10^{-6}$</td>
</tr>
<tr>
<td>$W$</td>
<td>$7.8 \times 10^{-6}$</td>
<td>$7.8 \times 10^{-6}$</td>
</tr>
<tr>
<td>$L_D$</td>
<td>$900.1 \times 10^{-12}$</td>
<td>$900.1 \times 10^{-12}$</td>
</tr>
<tr>
<td>$pV_{TH}, nV_{TH}$</td>
<td>$-0.8311$</td>
<td>$0.6081$</td>
</tr>
<tr>
<td>$k_p, k_n$</td>
<td>$19.34 \times 10^{-6}$</td>
<td>$74.21 \times 10^{-6}$</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>$0.1$</td>
<td>$0.2$</td>
</tr>
<tr>
<td>$p\gamma, n\gamma$</td>
<td>$0.3046$</td>
<td>$0.6166$</td>
</tr>
<tr>
<td>$C_{10}$</td>
<td>$259.97 \times 10^{-6}$</td>
<td>$280.65 \times 10^{-6}$</td>
</tr>
</tbody>
</table>
The gain is given by the following relation

\[ G_{out} = \frac{dV_{out}}{dE_1}. \]  

If the input voltage is set at series to \( E_1 \). It seems to have the maximum around \( E_1 = 0.37[V] \). Now, let us calculate the maximum gain with the steepest descent method shown is section 2.2. Our Spice-oriented optimization algorithm is realized the equivalent circuit shown by Fig.7, where the above 3 circuits are inverters, and the below 3 are corresponding sensitivity circuits, where they are set to \( E_1 = 1[V] \) in the sensitivity analysis because of the gain of \( E_1 \). The optimum point is obtained by the following steepest descent method shown Fig.4.

\[
\begin{align*}
\frac{dE_1}{dt} &= \frac{S_E(E_1)}{S_E(E_1)} S_{E_2}(E_1 + \Delta E_2) - S_{E_2}(E_1, E_2) \\
\frac{dE_2}{dt} &= -\frac{S_E(E_1, E_2)}{S_E(E_1)} S_{E_2}(E_1 + \Delta E_2) - S_{E_2}(E_1, E_2) 
\end{align*}
\]

Thus, the optimum point corresponds to the equilibrium point, which is found by the transient analysis of Spice. To investigate the result, we calculate the gain characteristic for the variable \( E_1 \) with \( E_2 = 1.946 \) as shown in Fig.8(b). The result is exactly equal to the result from Fig.8(a).

**AB amplifier**

Now, we consider AB amplifier shown in Fig.9(a) [1]. It is a class of differential amplifier and consisted of 14 MOS-FETs. At first, we optimize the gain by changing the sizes (\( nL, nW, pL, pW \)) between 1\( \mu m < L, W < 10\mu m \), where we classify the transistors into 3 groups according to pMOS mirror circuits (1,2,3,4,5,6), nMOS differential inputs and mirror circuit (1,3,4,5,6,8) and nMOS current sources (2,7). We also consider the bias voltage \( V_B \) so that we optimized seven parameters and have the result shown by Fig.9 (c).

5. Conclusions and remarks

In this paper, we propose a Spice-oriented design algorithm of CMOS amplifiers for attaining the maximum DC gain. The optimization algorithm is based on the steepest descent method such that the gradient direction is decided by the analysis of the sensitivity circuit. Since our optimization method is realized by the nonlinear RC circuits with ABMs, the solution is found by the transient analysis of Spice. From our numerical experiments, we can get the solution, stably. For future work, we will apply our method to bipolar transistor circuits and the others.

References

Abstract—A Verilog-AMS model for a fractional-$N$ frequency synthesizer is presented that is capable of predicting the noise and jitter performance. The model is based on a voltage-domain behavioral model. Compared to phase-domain models, voltage-domain models do a better job of capturing the details of the behavior of the loop, details such as signal capturing and escaping traces. Simulation efficiency is improved by merging the voltage controlled oscillator (VCO) and frequency divider. Due to the benefits of Verilog-AMS, the $\Delta \Sigma$ modulator which is incorporated in the synthesizer is modelled in the fully digital way. This makes it accurate enough to evaluate how the performance of the frequency synthesizer is related to the $\Delta \Sigma$ modulator. The spur-minimizing effect of an odd initial condition on the first accumulator of the $\Delta \Sigma$ modulator is verified in our simulation. The results are in agreement with prior published data on fractional-$N$ synthesizers.

I. INTRODUCTION

Phase-locked loop (PLL) based frequency synthesizers are widely used in all kinds of circuits. The fractional-$N$ approach to frequency synthesis enables fast dynamics to be achieved within the PLL by allowing a high reference frequency [1]. The fractional-$N$ divider is realized by changing the divider ratio between a range of integers to get an average fractional ratio. However, the changing of divider ratios in a periodic manner can produce excess spur and phase noise in the output of the frequency synthesizer. The $\Delta \Sigma$ modulation technique [1] is proposed to shape the noise produced by the dynamic divider ratio. This method has been widely used in a variety of applications ranging from accurate frequency generation to direct frequency modulation in transmitters [2].

Figure 1 illustrates such a PLL synthesizer where the output of the $\Delta \Sigma$ modulator is used to modulate the instantaneous division ratio of a multi–modulus frequency divider. The division ratio alterations take place very rapidly, and in a pseudo–random fashion so that the phase noise and spurious contents are pushed to higher frequencies, while the existing PLL loop filter can easily attenuate them. This technique permits narrow step sizes compared to the reference frequency and fast settling time, while improving the phase noise performance of the PLL frequency synthesizers [3].

All $\Delta \Sigma$ modulators produce quantization noise and make a considerable contribution to the overall noise of the fractional-$N$ synthesizer. However, the existing models for $\Delta \Sigma$ modulators have used C++ or Verilog-A [4–6], in which the $\Delta \Sigma$ modulator has been modeled using an idealized linear transfer function. Due to the benefit of Verilog-AMS, we can model the $\Delta \Sigma$ modulator based on its digital implementation. This makes it accurate enough to evaluate the nonlinear effects of the $\Delta \Sigma$ modulator. For example, the positive effect of a “1” LSB initial condition [7] is demonstrated in this paper.

The rest of the paper is organized as follows. In section II, $\Delta \Sigma$ modulator is modeled and explained. In section III, the full Verilog-AMS behavioral model for a $\Delta \Sigma$ fractional-$N$ frequency synthesizer is described. Simulation results are discussed in section IV and we summarize our conclusions in section V.

II. $\Delta \Sigma$ MODULATOR AND ITS MODELING

$\Delta \Sigma$ modulators are well known in the field of communications and have been used extensively in A/D and D/A conversion applications. The fundamental operation of these modulators relies on the fact that the spectrum of the quantization noise is shaped such that a small amount of noise power remains within the useful signal band with the rest of the quantization noise is pushed to higher frequencies.

The same principles can also be exploited in fractional-$N$ frequency synthesis applications by pushing the phase error toward the higher frequencies so that the phase noise in the vicinity of the desired carrier frequency is small. The high–frequency phase error is subsequently suppressed by the already existing loop filter, which has a low–pass characteristic.

The choice of an appropriate $\Delta \Sigma$ modulator structure for fractional-$N$ synthesis requires the consideration of many factors including noise shaping, spurious content of the output spectrum, output levels, loop filter order and circuit complexity [8]. In a digital implementation of the modulator,
an accumulator acts as an integrator and a comparator. By selecting higher order ∆Σ modulators, the spurious energy is spread out and shaped to high frequency, which is removed by the low-pass nature of the loop filter [8]. Higher order modulators can be realized with a MASH (Multi-state noise SHaping) architecture, in which a cascade of lower-order structures is used. It is typically constructed using a cascade of first order modulators.

In [7], it is shown that MASH ∆Σ modulators can provide a tone–free output spectrum when an “irrational” initial condition is used in the first accumulator. This can be accomplished by setting the LSB to “1” in the first accumulator each time the circuit is reset [7]. A digital implementation of the MASH 1-1 ∆Σ modulator with the “1” LSB initial condition in the first accumulator is depicted in Fig. 2. The effect of the “1” LSB initial condition related to the overall fractional-N frequency synthesizer can be observed in our simulation results which are shown in Sec. IV. The Verilog AMS realization of this kind of ∆Σ modulator is shown in Listing 1. Fig. 4 and Fig. 5 show the noise–shaping effect in the frequency domain for the ∆Σ modulator with and without the “1” initial condition, respectively. This is a second-order ∆Σ modulator, which means the spectrum of noise rises at 40 dB/dec, as expected.

III. MODEL DESCRIPTION

As regards the behavioral model for phase noise or jitter simulation, there are two different modeling methods. One is the popular linear phase-domain model [5], [9]. In the simplest case, these models are linear and analyzed easily in the frequency domain, making it simple to model and predict phase noise. The other approach is the strong nonlinear voltage–domain model, which we adopt in this paper; it formulates the models in terms of jitter [5]. This approach permits prediction of frequency synthesizer jitter behavior once the noise behavior of the blocks has been characterized. Generally, phase–domain models are considerably more efficient, but voltage–domain models do a better job of capturing the details of the behavior of the loop; details such as signal capturing and escaping traces [5].

Fast behavioral simulation of fractional-N synthesizers is challenging. The synthesizer’s high output frequency (often
in the gigahertz range) imposes a high simulation sample frequency for simulators. Unfortunately, the overall PLL dynamics have a bandwidth that is typically three to four orders of magnitude lower than the output frequency (often 100 kHz to 1 MHz of bandwidth compared to 1 to 10 GHz for the output frequency). Thus, simulators take a long time to compute the system’s dynamic response, because they must process many simulation samples. In the integer-$N$ PLL frequency synthesizer, using the property of the unchanged divide ratio, Kundert [5] proposed an approach to merging the VCO and frequency divider and thereby lowering the highest frequency of the system to that of the input reference frequency. However, the divider ratio is not fixed in fractional-$N$ frequency synthesizers. In order to overcome this problem, a solution is proposed in [6] for which the corresponding block diagram is shown in Fig. 6.

![Fig. 6. Improved model for fractional-$N$ PLL synthesizer with $\Delta\Sigma$ modulator](image)

If we consider a signal with frequency $f_1$ after an $N$ divider, the output signal will be at the frequency of $f_2 = f_1/N$. Since the divider outputs one pulse for every $N$ pulses at its input, the variance in the output period is the sum of the variance in $N$ input periods. Thus, from [5] we know that the jitter at the output is $\sqrt{N}$ times larger than the jitter at the input, or

$$J_2 = \sqrt{N}J_1. \quad (1)$$

We can see that if we want to get the jitter of a signal at a high frequency, we can observe the jitter of the signal after an ideal divider and then, using (1), we can get the noise level at the output of the VCO. The phase noise has the following relationship [5]:

$$S_{\phi 1} = N^2S_{\phi 2}. \quad (2)$$

In our model, we combine the VCO and an ideal fixed divider into a single block and divide its output by $1,\ldots, 1 + P/N$ which is controlled by the output of the $\Delta\Sigma$ modulator. $P$ is determined by the order of the $\Delta\Sigma$ modulator. Measuring the output signal of this block and using (2), we can obtain the noise level of the VCO output. The realization of this kind of merging in Verilog-AMS is shown in Listing 2.

IV. SIMULATION RESULTS AND DISCUSSION

We have applied these ideas to model and simulate a $\Delta\Sigma$ fractional-$N$ frequency synthesizer. The architecture and all the parameters are taken from [2]. The synthesizer is with a reference frequency of $f_{ref} = 20$ MHz, an output frequency of $f_{out} = 1846$ MHz, and a nominal divider ratio of $P = 92.3$. The bandwidth of the loop filter is $84$ kHz and the gain of the VCO is $30$ MHz/V. The charge pump current is $1.5\mu$A. The simulation is carried on a PC with a CPU of 2.4 GHz frequency and 512 MB memory. It takes 97 seconds to run the simulation for 24,000 reference cycles. The output data are transacted through the method proposed in [5]. The results are compensated for non-unity resolution bandwidth (-33 dB) and for the suppression of the dividers (39 dB). The simulation result for no “1” LSB irrational initial condition added is shown in Fig. 8, and Fig. 9 shows the result with a “1” LSB initial condition. As expected, the former shows much more noise resulting from the spurs of the $\Delta\Sigma$, while the later agrees qualitatively with the theoretical prediction in [2] which shown as the dashed curve. Note that the system without “1” LSB reset has significantly higher spurs than that with it. We observe that with a “1” LSB irrational initial condition, the power spectral density of the VCO phase can be suppressed by more than 10 dB out of the offset frequency of 1 MHz.

V. CONCLUSION

In this paper, a Verilog-AMS model for simulating the phase noise and jitter performance of $\Delta\Sigma$ fractional-$N$ frequency synthesizers is presented. In our implementation, the $\Delta\Sigma$ modulator is modeled through its digital implementation rather than the transfer function.

To our knowledge, this is the first complete nonlinear implementation. The power of this approach is that for the first time, the presence of coexisting attractors, has been demonstrated in a system level simulation. Until now, the effect of the initial condition has been masked by the linearized modeling approach to the $\Delta\Sigma$ modulator.

Furthermore, the combination of VCO and divider is adopted in our model to maximize the simulation efficiency. Our simulation results agree qualitatively with the theoretical prediction in [2]. Above 10 dB noise suppression is realized when adding a “1” initial condition.
ACKNOWLEDGMENT

This work is supported in part by Science Foundation Ireland under Grant 02/IN.1/I45.

REFERENCES


RLCG-MNA Formulation for Fast Simulation and MOR of RLC Networks

Yuichi Tanji*, Takayuki Watanabe**, and Hideki Asai***

*Dept. of RISE, Kagawa University, Takamatsu, 761-0396 Japan
**School of Administration & Informatics, University of Shizuoka, Shizuoka, 422-8526 Japan
***Dept. of Systems Eng. Shizuoka University, Hamamatsu, 432-8561 Japan

Abstract – RLCG-MNA formulation useful for both fast simulation and model order reduction (MOR) of linear RLC networks is presented. It is shown that the RLCG-MNA formulation makes the state-of-the-art simulation methods faster and improves accuracy of the MOR algorithms. Therefore, the RLCG-MNA formulation is effectively used for the time- and/or frequency-domain analysis which is indispensable for power and signal integrity estimations for high-speed and high-performance electronic systems. For some examples, the performance of the simulation and MOR methods for the RLC networks formulated by the RLCG-MNA equation is demonstrated.

I. Introduction

Increasing clock speed and low resistance metal on high performance integrated circuits make on-chip inductance effects prominent. Since the reactive component \(j\omega L\) of a conductor is comparable to the resistance \(R\) at high frequencies, the inductance gives rise to over and undershoot on propagation signals, which results in fault switching operations and large power dissipation. The complex 3D structural interconnects in integrated circuits are modeled by the RLC circuits equivalent to the electromagnetic fields. Since the RLC circuits are extremely large-scale, the time- and/or frequency-domain analysis for evaluating the inductance effects wastes huge CPU times. Moreover, the inductance issue is not restricted to VLSI systems. To ensure the reliability of overall performance of the electronic systems, the packages and PCBs embedded in the systems are modeled by RLC circuits and the time- and/or frequency-domain analysis must be carried out.

In this paper, the RLCG-MNA formulation is presented for the fast simulation and model order reduction (MOR) of the large scale RLC networks. The concept of the formulation is very simple. A pair of resistor and inductor which represents an impedance characteristic of conductor is treated as a branch, and KVL is applied to the branch without using KCL. The resistance matrix \(R\) results in the MNA matrix, although it does not exist in the original MNA equation which is widely used for circuit simulators. INDUCTWISE [1][2], the RPL algorithm [4], and LIM [5] are known as the state-of-art simulation methods, where the RLC networks are formulated by the MNA equation. Alternatively, formulating them by the RLCG-MNA formulation, we can accelerate INDUCTWISE and generalize the RLP algorithm and LIM. On the other hand, the MOR algorithms such as PRIMA [10] and SPRIM [11] reduce the order of equations, which are especially effective for the frequency-domain analysis. Then, we can improve accuracy of the MOR algorithms using the RLCG-MNA formulation.

For some examples, the performance of the simulation and MOR methods using the formulation is effectively demonstrated.

II. Fast Simulation

The fast simulation of RLC networks in the time-domain is achieved by formulating them by the RLCG-MNA formulation [6]-[9] and solving the set of equations with the concepts of nodal [1][2], mesh [4], and hybrid [5] analyses. The RLCG-MNA equation is written by

\[
\begin{align*}
\begin{bmatrix}
G & A^T \\
-A & R
\end{bmatrix}
\begin{bmatrix}
v(t) \\
i(t)
\end{bmatrix}
+ \begin{bmatrix}
C & 0 \\
0 & L
\end{bmatrix}
\frac{d}{dt}
\begin{bmatrix}
v(t) \\
i(t)
\end{bmatrix}
= \begin{bmatrix}
A_i^T \\
0
\end{bmatrix}
\begin{bmatrix}
I(t) \\
0
\end{bmatrix}.
\end{align*}
\]

(1)

where \(v(t)\) and \(i(t)\) are the node voltage and the inductor current vectors, respectively. \(G, C, R,\) and \(L\) are the conductance, capacitance, resistance, and inductance matrices, respectively. \(A_i^T\) is the transposition of incident matrix associated with the inductor branches. The current sources here are only assumed as the independent sources and are expressed by \(I(t)\). The dimensions of vectors and matrices are as follows,

\[
v(t) \in \mathbb{R}^k, i(t) \in \mathbb{R}^n, G, C \in \mathbb{R}^{k \times k}, R, L \in \mathbb{R}^{n \times n}.
\]

The RLCG-MNA formulation can be easily understood by considering the 2-conductor system shown in Fig. 1. The impedance matrix of the 2-conductor system is composed of two resistances, two self and one mutual inductances. There are two floating nodes which are connected with a resistor and an inductor. Not applying theses nodes to KCL, we apply KVL to the two branches consisting of a resistor and an inductor to obtain the RLCG-MNA equation (1). The KVL equations are expressed in the second low of (1). Since
the KCL equations related to the floating nodes connected with the resistor and inductor do not appear in (1), the RLCG-MNA equation is smaller than the original one.

Applying a numerical integration formula, we calculate the voltage and current vectors at a discrete point (we use the notations $v(t_i)=v_i$, $i(t_i)=I_i$, and $I(t_i)=I'$). In INDUCTWISE [1][2], the RLC networks are analyzed by the nodal analysis using the reactance matrix $L_i$. The locality of the reactance matrix [3] eases the cost of the sparse Cholesky decomposition and the preconditioning conjugate gradient method for solving the MNA equation at each time point. Based on the nodal analysis as INDUCTWISE, the voltage and current vectors of (1) are updated at a discrete time point as

$$G + \frac{1}{h} C + A' (R + \frac{1}{h} L) A) v^{k+1} = -G + \frac{1}{h} C - A' (R + \frac{1}{h} L) A) v^k - A' [I - (R + \frac{1}{h} L)] + A'(I^{k+1} + I^k),$$

$$i^{k+1} = (R + \frac{1}{h} L)^{-1} (R + \frac{1}{h} L) [R + \frac{1}{h} L]^{-1} (v^{k+1} + v^k).$$

(2)

The reference [5] shows that the RLP algorithm is several dozen times faster than INDUCTWISE. The RLP algorithm takes advantage of the inverse of the capacitance matrix $C_i$ rather than $L_i$ and captures the transient responses by the mesh analysis. Applying the mesh analysis to (1), we can calculate the responses as

$$v^{k+1} = \left(G + \frac{1}{h} C + A' (R + \frac{1}{h} L) A) v^k = \left(-G + \frac{1}{h} C - A' (R + \frac{1}{h} L) A) v^k - A' [I - (R + \frac{1}{h} L)] + A'(I^{k+1} + I^k),$$

$$i^{k+1} = \left(R + \frac{1}{h} L)^{-1} (R + \frac{1}{h} L) [R + \frac{1}{h} L]^{-1} (v^{k+1} + v^k).$$

(3)

In (4) and (5), the inverse of the nodal admittance matrix $G + \frac{1}{h} C$ is used instead of the inverse of the capacitance matrix in the RLP algorithm.

The third method is based on LIM [5] which is known as a circuit based FDTD method, where the voltage and current vectors are alternately updated. Although the system of equations related to either currents or voltages is solved in INDUCTWISE and the RLP algorithm, the system of equations associated with the current and voltage vectors is solved by LIM. Therefore, this method is a kind of hybrid analyses. Applying the hybrid analysis to (1), the update rule of the voltage and current is obtained from

$$\left(\frac{1}{2}G + \frac{1}{h} C \right)v^{k+1} = \left(\frac{1}{2}G - \frac{1}{h} C \right)v^k - A'T^{k+1/2} + A'I^{k+1/2},$$

$$\left(\frac{1}{2}R + \frac{1}{h} L \right)i^{k+1/2} = \left(\frac{1}{2}R - \frac{1}{h} L \right)i^{k-1/2} + Av^k.$$

If there does not exist a capacitor at a node or an inductor in a branch, a small value (latency) of capacitor or inductor must be assumed in LIM. By using (6) and (7), fortunately, the limitation is relaxed [8].

### III. MOR

The MNA equation must be solved at each frequency point in the frequency-domain analysis. Therefore, MOR prefers to the direct simulation, since the size of equation to be solved after MOR is much smaller than the direct case. Hence, we discuss the MOR of the RLCG-MNA equation (1) in this section.

For macromodeling, the circuit is to be expressed as a multiport network. The impedance matrix of an N-port network is written by

$$Z(s) = \begin{bmatrix} B^T & G & A' \\ 0 & -A & R \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix}^{-1} B.$$

(8)

where $B \in \mathbb{R}^{k \times N}$ is a selector matrix. Equation (8) is equivalent to

$$Z(s) = B^T(G + sC + A'(R + sL)^{-1}A)B.$$

(9)

In PRIMA [10], using the projection matrix $W$, the reduced-order model of (8) is obtained by

$$\tilde{Z}(s) = B^T W Z(s) W^T B.$$

(10)

In SPRIM [11], the projection matrix is reproduced as

$$\begin{bmatrix} W_1 & 0 \\ 0 & W_2 \end{bmatrix},$$

where $W_1 \in \mathbb{R}^{k \times q}$ and $W_2 \in \mathbb{R}^{n \times q}$ are parts of $W$ used in (10). Then, the reduced-order model is described by

$$\tilde{Z}(s) = B^T W \left[ \begin{bmatrix} G & A' \\ -A & R \end{bmatrix} + \begin{bmatrix} C & 0 \\ 0 & L \end{bmatrix} \right] W^T B.$$

(11)

where

$$G_i = W_i^T G W_1, A_i = W_i^T A W_1, C_i = W_i^T C W_1,$$

$$L_i = W_i^T L W_2, R_i = W_i^T R W_2.$$
The projection matrix is calculated by
\[
\begin{bmatrix}
W_1 \\
W_2
\end{bmatrix} = \text{colspan} \begin{bmatrix}
M_1^0 & M_1^1 & \cdots & M_1^{q-1}
\end{bmatrix}, \quad (12)
\]
where
\[
\begin{bmatrix}
M_1^0 \\
M_1^1
\end{bmatrix} \in \mathbb{R}^{k \times q}, M_2^j \in \mathbb{R}^{n \times q}
\]
is the \(j\)-th moment matrix which is a coefficient matrix of the Taylor expansion of RLCG-MNA equation in the frequency (Laplace) domain. The moment matrices are recursively calculated by
\[
\begin{bmatrix}
G + s_0 C + A^T (R + s_0 L)^{-1} A M_1^0 \\
- A R
\end{bmatrix} + s_0 \begin{bmatrix}
C & 0 \\
0 & L
\end{bmatrix} \begin{bmatrix}
M_1^j \\
M_2^j
\end{bmatrix} = \begin{bmatrix}
B \\
0
\end{bmatrix}, \quad (13)
\]
and (14) is solved by
\[
\begin{bmatrix}
G + s_0 C + A^T (R + s_0 L)^{-1} A M_1^j \\
- A R
\end{bmatrix} + s_0 \begin{bmatrix}
C & 0 \\
0 & L
\end{bmatrix} \begin{bmatrix}
M_1^{j-1} \\
M_2^{j-1}
\end{bmatrix} = \begin{bmatrix}
C & 0 \\
0 & L
\end{bmatrix} \begin{bmatrix}
M_1^{j-1} \\
M_2^{j-1}
\end{bmatrix} - \begin{bmatrix}
CM_1^{j-1} + A^T (R + s_0 L)^{-1} L M_2^{j-1}
\end{bmatrix}, \quad (15)
\]
Finally, the projection matrix for SPRIM is obtained from
\[
\begin{bmatrix}
\text{orth}(W_1) \\
0
\end{bmatrix} \quad \begin{bmatrix}
0 \\
\text{orth}(W_2)
\end{bmatrix}
\]
where \(\text{orth}(\cdot)\) means orthogonalization of columns of a matrix.

IV. Examples

The performance of both the simulation methods and the MOR algorithms incorporating the RLCG-MNA formulation was estimated by the power plane circuit shown in Fig. 2. The RLC network is one layer of a packaging system. All the computations were done on Matlab7 which was installed on Federa Core 3. The CPU times were measured on a Pentium 4 with 3 GHz clock and 1 Graduate memory.

A. Fast Simulation

The independent current source with a pulse waveform is given at the port1 and the transient waveforms are calculated by the nodal analysis (2) and (3), the mesh analysis (4) and (5), and the hybrid analysis (8) and (9). Table 1 in the next page shows the comparison of CPU times, where SCD means that (2) is solved by the sparse Cholesky decomposition, and PCG is the result by the preconditioning conjugate gradient method. Mesh and Hybrid mean the mesh and hybrid analyses, respectively. In all the simulations, sparsification is not carried out, thus, the results are almost same accuracy. We can see from Table 1 that the mesh analysis was the most inefficient. This is because the sparsification is not done though efficiency of the RLP algorithm greatly depends on the ratio of sparsified elements to the original ones. On the other hand, the hybrid analysis was the most efficient. In Table1, the results using the RLCG-MNA formulation are compared with the cases using the original MNA one. In SCD, the simulation with the RLCG-MNA formulation are about four times faster than with the original MNA one. In PCG, further speed-up is achieved. Since the application of the nodal analysis is not restricted by any circuit structure differently from the hybrid analysis, the nodal analysis of the RLCG-MNA is very attractive.

![Figure 2 Power plane circuit.](image)

<table>
<thead>
<tr>
<th>CPU times needed for MOR (MNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>cell</td>
</tr>
<tr>
<td>10×10</td>
</tr>
<tr>
<td>20×20</td>
</tr>
<tr>
<td>40×40</td>
</tr>
<tr>
<td>80×80</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CPU times needed for MOR (RLCG-MNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>cell</td>
</tr>
<tr>
<td>10×10</td>
</tr>
<tr>
<td>20×20</td>
</tr>
<tr>
<td>40×40</td>
</tr>
<tr>
<td>80×80</td>
</tr>
</tbody>
</table>

B. MOR

The driving point impedance at port1 in Fig. 3 is calculated by the MOR algorithms. Figures 3 and 4 show the error between the reduced-order model and exact one, where the order of \((G+sC+A^T(R+sL)^{-1}A)^{-1}\) in (9) is reduced from 441 to 30. In these figures, MNA1 is the result in which the
network is formulated by the original MNA equation and the moment matrices are calculated by (13) and (14) at $s_0=0$. MNA2 is the result in the case with $s_0=1$ and $R=0$ in (15) and (16) (we cannot take $s_0=0$ due to singularity). The network is formulated by the RLCG-MNA formulation in RLCG1 and RLCG2 and the moment matrices are obtained by (13) and (14) in RLCG1 and by (15) and (16) at $s_0=0$ in RLCG2.

We can see that when the moment matrices are calculated by (15) and (16), the frequency responses become less accurate than by (13) and (14) for both SRPM and SPRIM. However, the influence in the MOR using the original MNA formulation is larger than using the RLCG-MNA one. The results of MNA2 in Figs. 3 and 4 are less accurate than the others. On the other hand, the CPU times needed for the MOR’s are reduced by using (15) and (16) as shown in Tables 2 and 3, where “standard” means use of (13) and (14) and “economic” is (15) and (16).

**Table 1 CPU time comparison of (time-domain) simulation methods**

<table>
<thead>
<tr>
<th>Cell</th>
<th>element</th>
<th>SCD</th>
<th>PCG</th>
<th>Mesh</th>
<th>Hybrid</th>
</tr>
</thead>
<tbody>
<tr>
<td>10x10</td>
<td>561</td>
<td>2.02</td>
<td>17.44</td>
<td>0.96</td>
<td>4.2</td>
</tr>
<tr>
<td>20x20</td>
<td>2,121</td>
<td>9.8</td>
<td>52.5</td>
<td>3.1</td>
<td>2.43</td>
</tr>
<tr>
<td>40x40</td>
<td>8,241</td>
<td>61.73</td>
<td>215.42</td>
<td>19.47</td>
<td>8.23</td>
</tr>
<tr>
<td>80x80</td>
<td>32,481</td>
<td>411.08</td>
<td>924.47</td>
<td>125.53</td>
<td>30.69</td>
</tr>
<tr>
<td>160x160</td>
<td>128,961</td>
<td>2695.83</td>
<td>3688.55</td>
<td>885.54</td>
<td>149.47</td>
</tr>
</tbody>
</table>

V. Conclusions

The RLCG MNA formulation useful for both fast simulation and MOR of linear RLC networks has been presented, where it has been shown that this formulation makes the state-of-the-art simulation methods faster and improves accuracy of the MOR algorithms. Therefore, the MNA-RLCG formulation can be used in circuit simulators instead of the original MNA formulation.

References

Towards a quantitative theory of biocomputation

R. Stoop and T. Ott

Institute of Neuroinformatics ETHZ/UZH
Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: ruedi@ini.phys.ethz.ch

Abstract—The concepts of information, symbols, pattern, and computation are central to science. Although the most central among these notions, information and computation, have been cast mathematically, these definitions often generate difficulties when they are applied to biological systems. The common formal definition of ‘computation’ is by means of procedures consisting of well-defined steps, in an attempt to reach a well-defined goal (after reaching it, the machine halts). Many biological processes, however, do not seem to fit into this paradigm, mostly because they are based on poorly defined heuristics that are continuously re-weighted, interrupted, and otherwise altered in real-time. Generally, biological systems are open, because they exchange matter across their boundaries. This indicates that the usual understanding of the halting criterion as the hallmark of computation, may be too strict in this context.

1. Symbols and information

Usually, computation is defined in terms of ‘symbol manipulation’ or ‘symbol processing’. This embodies the problem that it is not at all obvious what a symbol is, implying that what it means to process or manipulate such a symbol, cannot be obvious as well. What is the nature of a symbol? How can a symbol or its emergence be defined in a constructive objective way? Knowledge that is mediated by human biology is not a human construct, but neither is a human-independent fact. It is probably more precise to see it as a projection of the ‘real world’ onto the sensory-temporal dimensions accessible by the humans. This, however, again is not the full truth, as this projection is not guided only by biology, but also by personal experience on the background of culture. Symbols worked out in this sense can be used to discover new symbols. In this way, the discovery of length leads to the discovery of depth and height, which is sufficient for the everyday practical purpose. As W. Heisenberg and E. Schrödinger put it: "We have to remember that what we observe is not nature itself, but nature exposed to our method of questioning", and "Every man’s world picture is and always remains a construct of his mind and cannot be proven to have any other existence", respectively. Science is the disciplinary approach to codify observations in symbols and their relations, so that they can pass some stringent tests for internal consistency. Note, however, that even the most successful truths (e.g., as Gödel proved, mathematics itself) can turn out to be incomplete in the sense to be applicable only to a subset (in particular to those who accept the symbols (like, e.g., infinity)). Distinction, grouping and naming of observations, are among the most basic cultural activities. More recently, notably in science, the focus has shifted from the identification of single objects, to the fusion of particular properties or objects towards bigger units (objects, classes of objects, respectively). This task of grouping together items that – in a situation-dependent sense – ‘belong together’ is referred to as clustering. Clustering has gained increased importance in all domains of data analysis and processing, the most prominent examples of which are bioinformatics, chemoinformatics, neuroscience, visual scene analysis or the design of autonomous systems. A clustering process, which is a second stage of processing, is supposed to work in a completely unbiased way, as in most situations, no apriori information about the number of clusters, their shapes, their internal structures (as the most relevant aspects), is available. If the clustering fails to detect satisfactory solutions, our conclusion will be that the first stage of the processing needs to be improved.

As a paradigmatic example, in the superparamagnetic clustering of N items with pairwise affinities d_{ij}, an inhomogeneous grid of Potts spins is constructed: Each item i is represented by one site of the grid with Potts spin variable s_i, where s_i \in \{1, ..., q\}. q is typically set to 10 or 20 \[1, 2\]. It is important to note that the choice of q is largely arbitrary and is not related to the number of emerging clusters. Each spin is symmetrically coupled to its k (not necessarily mutual) nearest neighbors. The choice of k will be discussed later on. The coupling strength J_{ij} is a decreasing function of d_{ij}, e.g. J_{ij} = J_{ii} = \frac{1}{K} \exp(-\frac{d_{ij}}{2K}). K is the average number of coupled neighbors per site (not necessarily equal to k). a is a local length scale which is set per default to the average distance between coupled spins. It will turn out, however, that sometimes it may be advantageous to adapt these length scales. Each spin configuration is characterized by an energy expressed by the Potts spin Hamiltonian H(s) = \sum_{i,j} J_{ij}(1-\delta_{s_{ij}}), where the sum runs over all connections (i,j) and s denotes a spin configuration. The system is considered in the canonical ensemble. The probability for a certain spin configuration is thus given by the Boltzmann distribution p(s) = \frac{1}{Z} e^{-H(s)/T}. As the temperature T is increased, inhomogeneous Potts
systems typically undergo a number of phase transitions. This is the inspirational source for the SC clustering algorithm. (I) For small $T$, the system is in the ferromagnetic phase, where spins like to be aligned. (II) At an intermediary $T$-range, a superparamagnetic phase occurs: Strongly coupled spins tend to be aligned, whereas weakly coupled spins behave independently. Thus, clusters of aligned spins occur, reflecting groups of similar data items. A further increase of $T$ generally leads to a continued breaking-up of these clusters into smaller clusters, so that a hierarchy of classes and subclasses is obtained. (III) For high $T$, the system enters the paramagnetic phase where any order disappears and only singleton clusters remain.

In the chemistry application of this approach, the rationale is given by the assumption that structurally similar compounds are likely to exhibit similar properties. The search for new drugs among thousands of different compounds in a chemical library can be greatly facilitated if the library can be sectioned into natural classes of structurally similar chemicals. For optimally chosen classes, efficient testing can be achieved. In [2], our SSC variant of SC was compared to the methods that in the field of chemoinformatics [3] are regarded to yield the most reliable results (notably Ward clustering). The test set comprised 153 compounds from 7 different chemical classes. Using chemical fingerprints to describe the compound’s structure, among the tested methods, SSC was the only one able to reliably reconstruct the seven involved chemical classes. To further elucidate this concept, we concentrate on a simple low-level task, the separation of objects (i.e., homogeneous regions) from a noisy background. Our working example is shown in Fig. 1, where two black objects of size 65 and 25 pixels, respectively, and three small white objects of size 12, 8, and 8 pixels, are embedded. For clustering, to each pixel $i$ a Potts spin $s_i$ is assigned. The connectivity graph is obtained by connecting each pixel to its 4 adjoined neighbors. The strength of a connection is again given by $J_{ij}$. The distance between two connected pixels $i$ and $j$ is determined by the pixel values difference, i.e., $d_{ij} = |f(i) - f(j)|$, where $f(i) = f(x_i, y_i)$ and the pixels are numbered in some way.

Given the symbols, the definition of computation is often given in terms of information: The number of distinctions necessary to specify some entity. This equates computation with information: A recognized pattern is precisely a bit of information. Relations between information are themselves informative, and can thereby serve as constraints on computational processes. Any aggregate of events or objects is said to contain ‘redundancy’ or ‘pattern’ if the aggregate can be divided in any way by a ‘slash mark’, such that an observer perceiving only what is on one side of the slash mark can guess, with better than random success, what is on the other side of the slash mark. In the language of cybernetics, the information available on one side of the slash will restrain (i.e., reduce the probability of) wrong guessing.

Information is computed in terms of entropy. As information can be encoded in many ways, the information content can be understood as the information content provided by the most efficient encoding. A faithful encoding of a pattern into a symbolic sequence has to obey, as the most stringent condition, the Kraft inequality. Among all the encodings that survive this condition, there are most efficient ones, those which equal their Huffman encoding. The entropy of such encodings can be considered as the information content of the process. The most often and most generally used definition of computation is that of ‘information processing’. In strict terms this implies that there is a well-defined mapping, from one set having a certain information to another set having some other information. Consequently, computation could possibly fall into one of two categories: 1) Information-preserving mappings: Information-preserving mappings lead to a reformulation of the problem, possibly and hopefully into a form that makes computation simpler. Rather than computation, these should therefore be considered as mappings between equivalent representations of the same information. 2) Information-reducing mappings: Among philosophers, such processes are also called synthesis or, owing to Wittgenstein, grammaticalization. During this process, information is decreased because fewer distinctions are made, but meaning is increased, because the smaller
number of distinctions has wider applicability. Compression mappings are ubiquitous in computation, and certainly constitute the majority of computations performed by living organisms. A simple example of this process is given by the dynamical system defined by the so-called roof map. Instead of describing its (chaotic) dynamics in terms of trajectories, a symbolic description is used. In this description, the symbols "0" and "1" distinguish between points x that are associated with slope one and two, respectively. The system can now be examined in the space of symbolic sequences. From the system, a nontrivial grammar emerges, in the sense that symbol "0" is always followed by symbol "1", but symbol "1" can be followed by either "0" or "1" (or: the symbol sequence "00" is prohibited). By the substitution of symbol "A" for sequence "01", a description of the system by means of a trivial grammar is obtained. It is easy to see that by modification of this example, systems with more involved (hierarchical) grammatical structures can be obtained. Grammaticalization yields a quick approximation to an object, by redefining a sequence or set of symbols in terms of a symbol. Thus, symbols may be defined as a processing hierarchy of processing hierarchies: a meta-hierarchy. A processing hierarchy (as opposed to a structural hierarchy) is a form of computation in which a slower computational process selects across (or – roughly equivalently – simplifies or compresses or analyzes or grammaticizes) a set of faster computational processes. This process of ‘grammaticalization’ selects computations that are consistent with respect to some ‘fitness function’ standing in for a goal or purpose.

2. A measure of computation

The most important practical problem in dealing with any system is to predict its evolution, i.e., the next state. In order to obtain the next value of a process, one needs to perform some experiment. It is the complexity of this experiment that should be taken as the definition of the complexity of the system. Consider a dynamical system with discrete time, defined by a map f on some set M in the Euclidean space \( \mathbb{R}^n \). Pick an arbitrary point \( x_0 \) in the phase space, take some neighborhood \( U = U(x_0) \) and consider the orbits \( \{f^n(U), n \in \mathbb{N}, U \} \). Take any observable \( \nu(x) \) and set \( \mu(x) = \exp(\nu(x)) \). We are interested in measures that are multiplicative along the orbit, i.e., for which the n-step average is evaluated as \( \left( \prod_{i=1}^{n} \nu(f^i(x)) \right)^{1/n} \). Examples of such observables are derivatives, probabilities, etc. Our goal is to study the problem of prediction of the next values \( \nu(f^{n+1}(x)) \) along the orbits and to evaluate the complexity of this prediction. We claim that this prediction problem is the one that defines an intrinsic complexity of dynamics. It is clear that practically any prediction should be probabilistic, because the exact value of the observable is not known (no measurement is precise) and because the outcome of a measurement is not constant in time.

For the decay of the probability of retaining a particular measurement value upon continued measurement, we employ the large deviation ansatz [4] \( P(\nu, n) \sim \exp\{-\epsilon \nu \} \). The thermodynamic formalism implies that \( g(\nu) = \nu - S(\nu) \), where \( S(\nu) \) is an entropy function. Requirements that apply to entropy functions are strict convexity with infinite derivatives at the two end-points of the curve (in the absence of phase transition effects). A conceptually simple observable is the finite-time Lyapunov exponent of n steps \( \epsilon(x, n) := -\frac{1}{n} \log | f^{(n)}(x) | \). The choice of this observable will simplify the illustration of the general construction of our measure. Our complexity measure, of the system plus observable, is defined as the complexity of prediction of the observable, averaged over all system behaviors captured by \( \epsilon \). The probability for observing trajectories with a specific value of \( \epsilon \), as a function of \( n \) behaves as

\[
P(\epsilon, n) \sim \exp\{-n(\epsilon - S(\epsilon))\}.
\]

The smaller \( \epsilon - S(\epsilon) \), the better the prediction based on the past of the orbits will be. Orbits with \( \epsilon = S(\epsilon) \) will yield perfect long-time prediction. Indeed, this value characterizes the natural invariant measure (e.g., [4]), providing the Lyapunov exponent. However, we seek a measure able to quantify the difficulty of making a correct prediction over all length scales. Suppose, for a thought experiment, that \( l > 1 \) invariant measures satisfy \( \epsilon_j = S(\epsilon_j), j = 1, \ldots, l \). Then, firstly, the system’s future obviously is more difficult to predict, and, secondly, this difficulty should be independent of the corresponding length scales \( \epsilon_j \). Moreover, the difficulty of making good predictions is also substantially increased by invariant measures for which \( \epsilon = S(\epsilon) \) is sufficiently close to zero (inversely, the larger this difference, to a lesser extent the corresponding invariant measures will contribute to the difficulty of making good predictions). The simplest candidate for a complexity measure taking these facts into account is the ratio \( S(\epsilon)/\epsilon \). To account for all system behaviors, the average \( \int S(\epsilon)/\epsilon \) d\( \epsilon \) is taken. In order to facilitate the comparison of systems with different topological entropies, we may rescale \( \epsilon \) and \( S(\epsilon) \) as \( \tilde{\epsilon} = \epsilon/\epsilon_0 \) and \( \tilde{S}(\tilde{\epsilon}) = S(\epsilon)/\epsilon_0 \), where \( \epsilon_0 \) is the topological length scale. With this, and in order to embrace repellors, we arrive at the most general form of the complexity measure

\[
C(\gamma, \beta)(\epsilon) = \epsilon_0^{2\beta} \int \frac{\epsilon_1}{\epsilon_1 - \kappa} \left( \frac{\tilde{S}(\tilde{\epsilon})}{\tilde{\epsilon}} \right)^\gamma d\tilde{\epsilon},
\]

where, to avoid divergence, we require \( -1 < \gamma \). The realization \( C(1, 1) \) is the natural measure of the difficulty of prediction, where the topological length scale is filtered out. This essentially rescales the time axes, and allows the comparison between systems with different time clocks (e.g., a ternary vs. a binary Bernoulli map). Although in most cases, the complexity measure needs to be calculated numerically, there are some analytical results. If the support of the specific entropy function only consists of one point (trivial spectrum), then complexity 0 is obtained. As a prominent example, zero complexity is obtained for the dynamically generated 1/3 Cantor set.
Consider now an arbitrary input, represented as the binary expansion of a number from the unit interval. This input can already be attributed a certain complexity (in the Kolmogorov sense, or more suitably in the context of measurements, in the sense of a complexity of prediction). By means of the system that performs the 'computation', this input complexity is transformed into an output complexity (measured, again, preferably as a complexity of prediction). The reduction of the input complexity by means of the map defines our measure of computation. With $C_s$ recognized as the appropriate measure of complexity, the most natural way to define computation is by means of the quantity

$$CO_s = \frac{1}{C_s(1,0) + 1}. \quad (3)$$

where the unity added prevents a possible singularity caused by $C_s(1,0) = 0$. This measure of computation is statistical in nature, as it is extracted by means of the thermodynamic formalism. Consequently, it is able to deal with real systems, whose precision is limited. It does not require explicit hierarchical analysis and is non-divergent by construction.

If the maximal scaling index $\nu_{\text{max}}$ is finite, the measure itself will be finite, bounded from below by zero and from above by the value $1 + \log(\nu_{\text{max}})$. $CO_s = 1$ indicates a system that performs optimal (decisive) computation, whereas $CO_s \leq \frac{1}{2}$ indicates that the system performs no notable computation. It is worth noting that in this way the measure of computation will be finite, in all realistic cases. Our measure of computation can be evaluated whenever for a system and observable the corresponding entropy function is known. These cases include experimental model cases as well as time series. The evaluation of the averages $CO_s$ is particularly simple for maps if a generating or approximated generating partition of the phase space, generated by the iteration of $f$, is available [5].

Seen from the computation they perform, the dynamical systems can then be divided into the following classes of increased computation (see Fig. 2):

$$CO_s(I) < 0.5 \leq CO_s(II) < CO_s(III) = 1. \quad (4)$$

Systems falling into the first class perform almost no computation. The most important example is provided by the intermittent systems, whose complexity when measured by means of $C_s(1,0)$, is highest among the prototypical dynamical systems [5]. This is in agreement with observations that biological complexity is highest at the border between order and chaos. Most standard dynamical systems belong to the second class. The third class is characterized by maximal computation $C_s(1,0) = 1$. Simple examples from this class are trivial maps (e.g., $f(x) \equiv 0$), the fully symmetric tent map, the Bernoulli shift map, and most fundamental arithmetic operations. To illustrate the latter claim, we may keep one operand fixed (e.g., to the value $x = a$) and take the map $f : y = x + a \mod 1$ as the representation of addition. Since the complexity of this map is zero, maximal computation $CO_s = 1$ is obtained. Similarly, multiplication by $a$ is represented by $f : y = ax \mod 1$.

![Figure 2: Filled circles: Averaged computation $CO_s$ for dynamical system families. 1) lowest computation (theoretical bound), 2) intermittent family (intermittency exponent $\gamma \in [1, 3]$), 3) tent map family (asymmetry $a \in [0, 1]$), 4) bungalow map family (right corner offset from center of interval $a \in [0, 0.5]$), 5) parabola family (nonlinearity parameter $a \in [a_\infty, 4]$, where $a_\infty$ is period doubling accumulation point), 6) maximal computation (trivial systems). Open circles: Results from a cat area V1 cortical neuron (left: suboptimal, right: optimal stimulation).](image)

References


Response of the Hopf cochlea to complex input sounds

S. Martignoli, A. Kern, and R. Stoop

Institute of Neuroinformatics ETHZ/UZH
Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: mstefan@ini.phys.ethz.ch, albert@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch

Abstract—Only recently, we are approaching a detailed mesoscopic understanding of the working principles of the mammalian hearing sensor, the cochlea. An important concept on this way is that of an active cochlear amplifier. Recently, a Hopf cochlea model has been implemented as an electronic circuit. The model consists of an array of sections each composed of a passive (hydrodynamic) and of an active (Hopf) amplificatory part, centered around logarithmically increasing characteristic frequencies. We show that for single-tone stimulation, response characteristics emerge that are close to those from a previous theoretical model and from the biological example. The effect of the Hopf bifurcation parameter upon the response is investigated in details. We then present measurements of the electronic cochlea’s response to multi-tones and clicks. Also for these signals, physiological basilar membrane measurements are available, which allows for a comparison of the hardware implementation with biology. We find that the cochlea’s response to these complex signals matches to the biological examples in a way that has not been achieved by earlier implementations.

1. Introduction

Understanding hearing in general and more specifically the mammalian hearing organ, the cochlea, is a long-standing human endeavor. The first basic step was taken by H. L. F. Helmholtz, who proposed in 1863 [1] the existence of a one-to-one correspondence between a position along the cochlear duct and a stimulation frequency to which, at the given position, the response is maximal (the tonotopic principle). Important steps that followed were von Békésy’s discovery of traveling waves along the basilar membrane (BM) as the carriers of the auditory information (1928 [2]) and Gold’s conjecture of an active amplification process within the cochlea (1948 [3]). The latter phenomenon was corroborated by the discovery of otoacoustic emissions (1978 [4]), the autonomous generation of sound by the cochlea itself. Subsequent experiments have revealed that the active amplification is located in the outer hair cells (OHC) attached to the basilar membrane [5, 6]. In turtles and bullfrogs, cells that are homologous to the mammalian hair cells [7] have been shown to display amplification profiles that are indicative of systems close to a Hopf instability [8].

Physiological measurements [11] reveal that Hopf-type amplification may also govern mammalian hearing. This led Eguíluz et al. [12] to propose to use Hopf-type amplifiers as the basic elements for cochlear modeling (for a related result, see [13]). They argued that the Hopf non-linearities could correctly capture the basic aspects of hearing: compression of the dynamic range; sharper tuning for lower sound intensities, and the generation of combination tones. Here, we demonstrate the electronic realization of this concept. From a detailed biophysical model of the cochlea, we derive the layout of a corresponding electronic realization. The responses generated by the electronic device are hardly distinguishable from biophysical measurements and validate in this way fully the Hopf amplifier approach.

In the cochlea, pressure variations generated by incoming sounds are transformed into incompressible and inviscid hydrodynamic waves. As these waves move down the cochlea, they cause small basilar membrane (BM) displacements [14]. Using \( x \) to denote the distance from the steps along the unrolled cochlea, the system can be linearly described by a water-surface wave with fluid depth \( h \), density \( \rho \), surface mass density \( m \) and exponentially decreasing transversal stiffness \( E(x) = E_0e^{-\alpha x} \) [15]. In this description, via the dispersion relation, the wave group velocity \( v_G \) and the wave number \( k \) are related by

\[
v_G = \frac{k}{E} = \frac{1}{m} \sinh(kh) = \frac{1}{m} \cosh^{-1}\left(\frac{c}{G_{\text{in}}(0)}\right)\]  

[16]. BM locations \( x \) and \( x_e(\omega) \) respond maximally by (passive) displacements when stimulated at characteristic frequencies

\[
\omega_e(x) = \sqrt{E(x)/m}. \tag{1}
\]

This relationship defines the tonotopic map. It can be shown that \( k(x, \omega) \) diverges as \( \omega \) approaches \( \omega_e(x) \) and that, as \( x \) approaches \( x_e(\omega) \) for fixed \( \omega \), the traveling wave stalls \((v_G = 0)\) at the point of (passive) resonance. Due to dissipative losses, the wave amplitude will reach a maximum at \( x < x_e(\omega) \) [17]. From the energy balance equation

\[
\frac{\partial e}{\partial t} = -\frac{1}{v_G(x, \omega)} \left[ \frac{\partial v_G(x, \omega)}{\partial x} + d(x, \omega) \right] e + \frac{a(x, e, \omega)}{v_G(x, \omega)}. \tag{2}
\]

In this equation, the local energy \( a(\cdot) \) supplied by the active amplification, works against the internal viscous losses

\[
d(x, \omega) = 4\nu k(x, \omega)^2, \quad \nu \text{ the kinematic viscosity} \ [18, 19].
\]
Following the biological example, the active amplification results from an array of Hopf-type power sources aligned along the BM, where each amplifier has its own natural frequency $\omega_{ch}(x)$. Given a forcing frequency $\omega$, the Hopf amplifiers with $\omega_{ch}(x) \approx \omega$ are maximally excited at locations $x_{ch}(\omega) < x_{ch}(\omega')$, before viscosity leads to a precipitous decay of the wave amplitude [17].

The characteristics of the active contribution $a$ are derived from a $\omega_{ch}$-rescaled Hopf differential equation

$$\dot{z} = (\mu + j) \omega_{ch} z - \omega_{ch} |z|^2 z - \omega_{ch} F(t), \quad z \in \mathbb{C}, \quad (3)$$

where $j$ is the imaginary unit. Assuming a 1:1 locking between signal and system, $z(t) = Re^{j\omega_{ch}t}$ is the amplified external periodic input $F = Fe^{j\omega t}$, $\omega_{ch}$ is the natural frequency of the oscillation, and $\mu \in \mathbb{R}$ denotes the Hopf nonlinearity parameter. For $F = 0$, Eq. (3) describes the generic differential equation displaying a Hopf bifurcation: For $\mu < 0$, the solution $z(t) = 0$ is a stable fixed point, whereas for $\mu > 0$, the fixed-point solution becomes unstable and a stable limit-cycle of the form $z(t) = \sqrt{\mu} e^{j\omega_{ch}t}$ appears. For time-varying signals $F(t)$ it is convenient to put a handle on the response latency, by multiplying the damping term $\omega_{ch} |z|^2 z$ by a factor $\gamma$. A nonzero forcing $F$ then yields $\omega_{ch} F e^{-\gamma t} = (\mu + j) \omega_{ch} R e^{j\omega t} - j\mu R$. Evaluation of the squared modulus and introducing the variable $\phi = \frac{\omega_{ch}}{\mu}$ results in

$$F^2 = \gamma^2 R^2 - 2\mu R^3 + \left[\mu^2 + (1 - \phi)^2\right] R^2, \quad (4)$$

which is easily solved. For $\mu = 0$ and close to resonance ($\omega = \omega_{ch}$), the response $R \approx F^{1/3}$ emerges, which forces the gain $G = R/F = F^{-2/3}$ to increase towards infinity as $F$ approaches zero. For $\mu < 0$, maintaining $\omega = \omega_{ch}$, we obtain the response $R = -F/\mu$ for weak stimuli $F$. As $F$ increases, the term $R^3$ in Eq. (4) can no longer be neglected, and the compressive nonlinear regime is entered ($\gamma R^2 \approx \mu^2 R^3 + 2\mu R^2$). The transition point is located at $\gamma F_c(\mu) \approx 0.91(-\mu)^{3/2}/\gamma^2$. For weak stimuli $F$, the response $R$ is almost linear; for moderate stimuli the differential gain of the system, $dR/dF$, decreases with increasing stimulus intensity. As we move away from resonance, the last term of Eq. (4) dominates, leading, as $R \approx F/|1 - \phi|$, to a linear response, irrespective of the stimulation strength. As the bandwidth $\Gamma \sim |\mu|/F$ for $F \leq F_C$ (and $\Gamma \sim \gamma^2 F^{2/3}$ for $F > F_C$), small $|\mu|$-values act as a high Q-factors (sharp resonances).

Among all the bifurcations that could possibly serve as the basis for an amplificatory system [9, 10], exactly the Hopf bifurcation provides the correct amplification laws that are observed in the mammalian ear (see Fig. 1a, b). The response $R = F^{1/3}$ in the neighborhood of the resonance frequency, which forces the gain $G = R/F = F^{-2/3}$ to increase towards infinity as $F$ approaches zero, provides the incredible sensitivity of the ear towards faint sounds. As $F$ increases, we enter a compressive nonlinear regime, which expresses our ability of coping with very loud sounds. The $\mu > 0$ regime is of no direct importance in the present context. The stable limit-cycles that emerge under this condition may, however, be important for the description of the otoacoustic emissions. For an intuitive understanding, at a nonzero stimulation, the Hopf amplifier can be interpreted as a nonlinear filter characterized by a tunable gain control (“quality factor”) $\mu$ (the Hopf bifurcation parameter) and an envelope detector $|z|^2$, where $z$ is the signal that is usually written in complex notation.

**Figure 1:** Comparison of response and gain, respectively, upon sustained single frequency stimulation, between a), b) biology [21] and c), d) hardware. $\mu = -0.1$.

### 2. Electronic sensor design

For stationary inputs, a differential equation based on energy balance arguments can be derived, which provides responses that are close to the physiological examples [17]. For transient signals we are, however, confronted with excessive computational demand. In order to arrive at a feasible hearing sensor, we resorted to a hardware implementation, for which we took care to preserve the insights gained from the theoretical model. We decomposed the cochlea into $n$ sections of characteristic frequencies $\omega_i$, $i = 1, \ldots, n$, and endowed each one with properties of the passive hydrodynamic behavior and an active Hopf amplifier, following closely the biological example. Major challenges were to properly connect the passive/active components towards a section, and the sections towards a cascade representing the entire cochlea.

The active amplification part can be approximated in circuitry by using a combination of integrative summers and multipliers, resistors $R$ and a capacitance $C$ [22]. The responses generated by the biophysically detailed model [17]...
suggest that the passive part can be modeled as a (6th-order) Butterworth circuit. The cochlea was constructed by connecting sections of logarithmically decreasing center frequencies $\omega_{i1}$ in series. In order to build a generic section, it was sufficient to specify the detuning between the passive frequency $\omega_i$ and the Hopf amplifier frequency $\omega_{iH}$, as $\phi_i(i) := \omega_{i1}/\omega_{iH} < 1$, $i = 1...n$, where $n$ is the number of sections. This was our first design parameter. The second design parameter was the relationship between the characteristic frequencies of subsequent sections $\Psi(i) = \frac{\omega_{i1}}{\omega_{i2}}$, $i = 1...n - 1$.

3. Measurements

For the realization of a hardware cochlea composed of 5 sections spanning one octave in the speech regime, we chose for simplicity the design parameters independent from the section, as $\Psi(i) = \Psi = 0.84$, $\phi_i(i) = \phi_s = 1.05^{-\frac{i}{2}}$, $\forall i$ (for a general frequency range to be covered, we use the second design parameter to determine the number of composing sections and their characteristic frequencies). In order to study the effects of the discretization of the cochlea, we varied the measurement point along the cochlea and the tuning width. The tuning width is expressed by the gain control $v_p$, which is the electronic equivalent of the Hopf parameter $\mu$. From this hardware cochlea, the amplitude $Amp = |v_o|$ generated in response to varying input frequencies and stimulation strengths was measured. The comparison of the measurements after the first and the second section demonstrates that by passing through the sections, the signal is gradually shaped (Fig. 3). After a few sections, the response attains the characteristic form that faithfully reproduces the results of the biophysical example.

By the electronic device, two-tone suppression and combination-tone generation, as the salient nonlinear phenomena of mammalian hearing, are reliably reproduced, see Fig. 4 and Fig. 5.

The response to transient, broadband stimulation is of similar quality. In Fig. 6, the results for click stimulations are displayed. A comparison with the corresponding biophysical measurements [26] demonstrates the close match. Also other characteristic features of the biophysical experiments of Ref. [26], as e.g., the instantaneous frequency response, are reliably reproduced.

4. Conclusions

The Hopf cochlear amplifier concept together with the cochlear biophysics thus provides the design of an extremely sensitive and robust hearing sensor, where
salient nonlinear signal processing characteristics (compressive nonlinearity, high sensitivity, two-tone suppression, combination-tone generation) are naturally implemented. By following closely the biological example, it is in the nature of the design of our cochlea that the response could be optimized by a response-dependent value of $\mu$, potentially implemented on the level of hair cells with its channel dynamics (small $|\mu|$ at signal onset, larger value with increased response). Even more importantly, on slower time-scale, the auditory neuronal feedback loop might tune $\mu$ in order to suppress unwanted, and to enhance desired, signal components. Experiments exploiting these features are currently under way.

Helpful collaboration with W. Schwarz is gratefully acknowledged.

References


Acoustic Source Separation by Atomic Signal Decomposition

A. Kern, Th. Ott, and R. Stoop

Institute of Neuroinformatics, University / ETH Zürich, 8057 Zürich
e-mail: albert@ini.phys.ethz.ch, WWW: http://www.stoop.net/group

Abstract—The separation of a mixed acoustic signal into its constituent sources is still a hard and challenging problem. We propose an approach to speech signal separation, which is based on a sparse time-frequency representation by local cosine packets. The separation is obtained by grouping (or: clustering) the individual local cosine packets to sets corresponding to the different sources. This grouping is performed by employing a criterion involving only the local frequency structure of the speech signal. In this way, it is possible to separate voiced speech segments. The grouping criterion can, however, be readily extended to include other, eg. time-related, speech features, such that successful speaker separation can be achieved in the general case. The sparse representation used leads to a very noise-robust separation method, which is crucial for real-world applications.

1. Introduction

The separation of a mixture of speech signals (cocktail party problem) is very challenging, with a broad range of possible applications. Besides computer-based speech recognition systems, the treatment of hearing damage by hearing aids will profit from a solution. While the healthy human auditory system manages to separate and associate different acoustic signals to their sources (auditory scene analysis [1]), hearing impaired people often have great difficulties with this task, even if endowed with sophisticated hearing aids. It would therefore be a great benefit if hearing aids would allow to extract a specific sound source.

In the last two decades, blind source separation methods have been developed in order to tackle the cocktail party problem. Efficient separation results may be obtained by statistical methods like ICA (independent component analysis) under rather restrictive conditions [2]. These include the presence of at least as many channels (microphones) as sound sources, the absence of background noise, and instantaneous mixing of the sources (ie. no time delays). Therefore, in realistic situations no satisfactory solution of the cocktail party problem can generally be obtained, even when using more sophisticated methods like noisy or convolutive ICA [3]. Furthermore, if only one microphone is available, the ICA method has to be abandoned altogether.

In this contribution, we propose an approach to single channel speech signal separation, by which the limitations posed by traditional ICA may be overcome. The only assumption made by ICA is that the recovered sources are independent realizations of a random process. Furthermore, it is also assumed that the sources do not exhibit any structure in time. However, it is precisely the distinct time-frequency structure of speech signals, that the auditory system uses for resolving the source separation problem [1].

In order to attain efficient acoustic source separation in natural environments, it is thus beneficial to encode such information in a mathematical framework. We will demonstrate that such a framework is provided by decomposing the speech signal into a superposition of time-frequency “atoms”, that are chosen from an appropriate dictionary. In this way, a sparse representation of the signal is obtained, which allows us to efficiently remove background noise, which is not restricted to be stationary nor Gaussian.

Source separation is obtained by grouping (or: clustering) the individual signal components (which are determined by the sparse representation) to sets corresponding to the different sources. This grouping is performed by employing a criterion involving only the local frequency structure of the speech signal. In this way, it is possible to separate voiced speech segments. The grouping criterion can, however, be readily extended to include other speech features, such that successful speaker separation can be achieved in more general situations.

2. Atomic Decomposition

Suppose that we are given a signal \( f(t) \in L^2(\mathbb{R}) \). Let us assume further that we are given a dictionary \( \mathcal{D} \), from which “time-frequency atoms” \( \phi_{\gamma} \) are chosen (\( \gamma \in \Gamma \) denotes a possibly multi-dimensional index). The representation

\[
f(t) = \sum_{\gamma} \alpha_{\gamma} \phi_{\gamma}, \quad \phi_{\gamma} \in \mathcal{D}
\]

is then termed an atomic decomposition of \( f(t) \), where we denote by \( \gamma \) the selected indices \( \gamma \in \Gamma \). The dictionary \( \mathcal{D} \) may constitute a basis of \( L^2(\mathbb{R}) \), or, preferentially, it may be overcomplete. In the latter case, several subsets of \( \mathcal{D} \) constitute a basis for \( L^2(\mathbb{R}) \), so that the representation (1) is not unique (ie. different sets \( \{\gamma_i\} \) may be selected). This flexibility may allow us to select the atoms \( \phi_{\gamma_i} \) in such a way that the decomposition (1) involves only a small number of terms, so that a sparse representation of the signal \( f(t) \) is obtained. To find a sparse representation of \( f(t) \) (given a dictionary \( \mathcal{D} \)) is a nontrivial problem. In the last decade, different algorithms have been proposed for obtaining approximate solutions.
3. Sparse Signal Representation

A sparse representation of \( f(t) \) is obtained if the \( \ell_0 \)-norm of the coefficient vector \( \alpha = (\alpha_n)_{n=1,\ldots}\infty \) is minimized,

\[
\min \|\alpha\|_0 \quad \text{such that} \quad f = \sum_{n} \alpha_n \phi_n. \tag{2}
\]

The \( \ell_0 \)-norm of a vector is given by the number of its non-zero elements. Unfortunately, (2) constitutes an NP-hard problem [4]. A feasible algorithm providing a solution to the sparse representation problem can thus only be realized if the condition (2) is relaxed. This may be done by replacing the \( \ell_0 \)-norm by the \( \ell_1 \)-norm, \( \|\alpha\|_1 = \sum_{i} |\alpha_i| \). It has recently been shown that under certain conditions \( \ell_0 \)- and \( \ell_1 \)-minimization are equivalent [5, 6].

The minimization of the \( \ell_1 \)-norm is achieved by the basis pursuit method [4] and constitutes a computationally demanding convex optimization problem. Basis pursuit starts from an initial representation (1), which is then optimized in respect to \( \|\alpha\|_1 \) by substituting atoms by more useful ones. Replacement of this global optimization problem by a stepwise decomposition approach significantly increases computational efficiency; in this case, however, since it is no longer guaranteed that the (relaxed) sparseness condition is satisfied, only approximate sparse representations will be obtained.

An efficient algorithm for stepwise generating an approximate sparse representation, called matching pursuit, was originally proposed by Mallat and Zhang [7]. Due to its computational efficiency, we will adopt this approach.

4. Matching Pursuit

The matching pursuit algorithm starts with \( f^{(0)} = 0 \) for the initial sparse approximation of the signal \( f \), and \( R^{(0)} = f \) for the approximation error (residuum). At step \( m \), matching pursuit provides the decomposition

\[
f = \sum_{i=0}^{m-1} \langle R^i f, \phi_{n_i} \rangle \phi_{n_i} + R^m f, \tag{3}
\]

\[
f^{(m)} = \sum_{i=0}^{m} \langle R^i f, \phi_{n_i} \rangle \phi_{n_i}, \tag{4}
\]

where \( f^{(m)} \) denotes the sparse approximation to \( f(t) \) at step \( m \), \( R^m f = f - f^{(m)} \) denotes the residual at step \( m \), and \( \langle \cdot, \cdot \rangle \) is the scalar product. From (3), it follows that \( f^{(m)} \in V^m = \text{span} \{ \phi_{n_0}, \ldots, \phi_{n_{m-1}} \} \), where \( V^m \) denotes the approximation space at step \( m \).

At each step \( k \), the atom \( \phi_{n_k} \in D \) is chosen according to the selection criterion

\[
|\alpha_k| = | \langle R^{k-1} f, \phi_{n_k} \rangle | = \max.
\]

(5)

Since always the atom whose interaction with the residuum \( R^{k-1} f \) is largest is selected, matching pursuit is sometimes also referred to as a “greedy” algorithm. The procedure may either be stopped after \( m \) steps (where \( m \) is determined in advance), or if \( \|R^m f\| \) falls below a predetermined threshold.

The described algorithm is optimal for orthogonal dictionaries \( D \); if \( D \) is non-orthogonal (which is the case if \( D \) is overcomplete) unsatisfactory results will occur, since \( \langle \phi_{n_i}, R^m f \rangle \neq 0 \) for some \( i \leq m \). This problem is alleviated by an additional back-projection step [7].

Since the matching pursuit algorithm is myopic (it generates the approximation (1) step by step), it may choose wrongly in the first few iterations. As a consequence, the resulting error must be corrected in many subsequent iterations, resulting in a non-sparse signal decomposition. For resolving this problem, sophisticated methods like optimized orthogonal matching pursuit (OOMP) [8] and backward-optimized orthogonal matching pursuit (BOOMP) [9] have been devised.

5. Local Cosine Packet Dictionaries

The choice of an appropriate dictionary \( D \) determines how sparse the obtained decomposition (1) may be. For signals \( f(t) \) exhibiting scaling properties, an overcomplete wavelet packet dictionary provides a sparse and computationally efficient representation. However, since speech signals have no underlying scaling structure, a wavelet-based dictionary proves less valuable: The formant structure of voiced speech segments demands a fine frequency resolution, even at high frequencies, while the resolution of the wavelet decomposition rapidly decays at higher frequencies. Although wavelet packets may be useful for determining rapid transitions (which are, e.g., expressed by plosive consonants like /bl, tl/), we therefore propose to use another dictionary which is better adapted to the properties of speech signals. Such a dictionary, which still allows for efficient computation, is provided by local cosine packets (LCP).

Local cosine packets are functions with finite time support, while being located (within the limits of Heisenberg uncertainty) in frequency,

\[
\phi_{i,j,k}(t) = \sqrt{\frac{2}{L_j}} b_j(t) \cos \frac{\pi}{L_j} \left( k + \frac{1}{2} \right) (t - a_{i,j}). \tag{6}
\]

\( b_j(t) \) is a smooth window function (“bell”) of length \( L_j = a_{i+1,j} - a_{i,j} \), ensuring that the LCP \( \phi_{i,j,k}(t) \) has time support in \([a_{i,j} - \epsilon, a_{i+1,j} + \epsilon]\) (\( \epsilon \) depending on the tapering of \( b_j(t) \)). The frequency pertaining to \( \phi_{i,j,k}(t) \) is determined by

\[
f_{k,j} = \frac{1}{2L_j} \left( k + \frac{1}{2} \right). \tag{7}
\]

A fast matching pursuit algorithm exists for dyadic LCP, where \( L_j = N/2^j \) (\( N \) being the signal length) and \( a_{i,j} = iL_j \).

According to (6), LCP are suitable if the signal energy is concentrated on relatively slowly varying frequency bands. This is the case for voiced speech segments. However,
speech signals also contain strongly modulated segments and fast transitions (at vowel on-/offsets and near consonants). In these situations, the matching pursuit algorithm employs a large number of LCP atoms with low coefficient values, which may result in a non-sparse representation. This deficiency can be remedied by either resorting to the basis pursuit method (ie., by relying on global optimization) or by using alternative dictionaries which allow to better match rapid frequency modulations (like chirplet dictionaries [10]) – but both at the expense of increased computational demand.

6. Noise Cleaning

Let us consider a speech signal $f(t)$ that is contaminated by (not necessarily stationary or Gaussian) noise $n(t)$, resulting in the detected signal

$$s(t) = f(t) + n(t).$$

Provided that the dictionary $D$ captures the structure of the signal $f(t)$, some atoms $\phi_{m}$ will interact strongest with the signal $f(t)$,

$$\langle \phi_{m}, f \rangle \gg \langle \phi_{m}, n \rangle \quad \forall j \neq i.$$  

The largest coefficients $\alpha_{ni} = \langle \phi_{ni}, R^{1} s \rangle \approx \langle \phi_{ni}, R^{1} f \rangle$ are thus expected to pertain to the signal $f(t)$,

$$f_{\text{clean}} = \sum_{i=0}^{m-1} \langle R^{1} s, \phi_{ni} \rangle \phi_{ni}.$$  

Truncation of the decomposition obtained by the matching pursuit algorithm therefore results in efficient noise cleaning.

Fig. 1 depicts the noise cleaning results for the situation where a speech sample has been contaminated by Gaussian noise bursts. Efficient noise cleaning has also been achieved for non-Gaussian $\alpha$-stable noise (which is very difficult to clean by traditional methods) and noise that is filtered by time-varying bandpass filters.

7. Separation of Voiced Speech Signals

7.1. Separation Method

We attempt to separate two signals (referenced by the index $s \in \{1, 2\}$) from a mixture. The separation procedure we use consists of the following steps: 1) The (usually time-varying) fundamental frequencies $f_{s}(t)$ pertaining to the signal $s$ need to be determined. If the modulation rates are not too fast, $f_{s}(t)$ ($s = 1, 2$) correspond to the two lowest incommensurable frequencies of the LCP having support at time $t$. 2) It has to be determined whether a selected LCP $\phi_{i,j,k}$ belongs to signal $s$. This is achieved by constructing an estimator $d_{i,j,k}^{s}$ if $d_{i,j,k}^{s} = \alpha d_{i,j,k}^{s'}$, where $\alpha > 1$ and $s' \neq s$ denotes the other signal, then $\phi_{i,j,k}$ is assigned to signal $s$, and vice versa. If no assignment can be made, then $\phi_{i,j,k}$ is considered as pertaining to background noise and is omitted.

The definition of the estimator $d_{i,j,k}^{s}$ relies on the specific time-frequency structure of voiced speech: Voiced speech consists of frequency bands whose frequencies are multiples of a fundamental frequency; in terms of the LCP decomposition this means that the frequencies $f_{a}$ and $f_{b}$ of two LCP belonging to the voiced speech segment of the same signal are commensurable, ie. $m f_{a} = n f_{b}$ for $m, n \in \mathbb{N}$.

First, the set $S_{i,j,k}$ of all LCP occurring synchronously with the selected LCP $\phi_{i,j,k}$ is determined, where we consider two LCP as synchronous if a significant portion of their time supports overlaps. Then, the differences between the frequency $f$ of $\phi_{i,j,k}$ and the frequencies of the LCP contained in the set $S_{i,j,k}$ are determined. Let us denote these differences by $\delta f_{r}$, where $r \in \{1, \ldots, N\}$ and $N$ is the cardinality of $S_{i,j,k}$. If $\phi_{i,j,k}$ belongs to signal $s$, due to the time-frequency structure of voiced speech, $\delta f_{r} = m f_{a}$ (for some $m \in \mathbb{N}$) holds for a significant number of indices $r$. Thus, the histogram of the fractional part $\rho_{r}$ of $\delta f_{r} / f_{a}$ will be peaked around zero\(^{3}\). On the other hand, if the LCP

\(^{3}\)Note that $\rho_{r} \in [0, 1]$. Since the situations $\rho_{r} = 1$ and $\rho_{r} = 0$ are identical, we re-map $\rho_{r}$ to the interval $[-0.5, 0.5]$. 

1021
and a number \( b > d \).

\[
\phi_{i,j,k} \text{ does not belong to signal } s, \text{ we expect a flat distribution of } \rho_r.
\]

These observations motivate several definitions of the estimator \( d_{i,j,k}^s \), which all lead to similar separation results. For example, we select a small interval \([-a, a]\) around zero and a number \( b > 0 \). Then, we compared the number \( N_a \) of indices \( r \) for which \( \rho_r \in [-a, a] \) to the number \( N_b \) of indices \( r \) for which \( \rho_r > b \). This leads to the estimator \( d_{i,j,k}^s = N_a / N_b \). Once the parameters \( a \) and \( b \) have been optimized, they are kept fixed for all signal separation problems; thus, no a priori signal information is needed.

### 7.2. Separation of Voiced Speech

By relying on the method outlined above, we managed to separate (1) two simultaneously uttered vowels (\( /a/ \) (black) and \( /i/ \) (red) and (b) vowel \( /a/ \) (red) and “arigato” (black). Bottom row: Separation results of (c) vowels \( /a/ \) and \( /i/ \) and (d) \( /a/ \) and “arigato”.

\[
\text{Figure 2: Separation of voiced speech segments. The supports of LCP in normalized time and frequency are shown. Top row: Superimposed LCP representations of (a) vowels } /a/ \text{ (black) and } /i/ \text{ (red) and (b) vowel } /a/ \text{ (red) and “arigato” (black). Bottom row: Separation results of (c) vowels } /a/ \text{ and } /i/ \text{ and (d) } /a/ \text{ and “arigato”.}
\]

8. Conclusion

Based on a sparse time-frequency representation of speech signals, convincing and, in contrast to ICA-based methods, noise-robust source separation results are obtained for voiced speech signals. This is a first step towards the separation of general speech mixtures, which also includes fast modulations, transients and unvoiced segments. Some improvements will be achieved by extending the LCP dictionary by other atoms (eg. using a chirplet dictionary [10]), for taking modulations and fast transients into account. For a general solution to the problem, we propose to combine our approach with statistical and information-theoretic methods like ICA: In a first step, preliminary estimates for the sources are obtained by our method, where as much a-priori information on the speech structure as possible is taken into account. In a next step, these estimates are optimized, eg. by maximizing the mutual independence between the sources (ie. ensuring small mutual information), while at the same time the correlation along a source estimate is kept large.

### References

Loopy belief propagation: Benefits and pitfalls on Ising-like systems

N. Stoop, T. Ott, and R. Stoop

Institute of Neuroinformatics ETHZ/UZH
Winterthurerstr. 190, 8057 Zürich, Switzerland
Email: norbert@ini.phys.ethz.ch, tott@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch

Abstract—Belief propagation (BeP) is a candidate for the accelerated evaluation of statistical averages if compared to Monte-Carlo approaches. For binary systems on infinite grids or with periodic boundary conditions (regular grids), we investigate the physical fixed points and their stability. Critical slowing down of the method is observed at the 2nd-order phase transition with \( T_C \approx 2.89 \). Above \( T_C \), convergence is guaranteed. Below the critical temperature \( T_C \), BeP convergence depends dramatically on the choice of initial conditions. This leads to convergence patterns typical for fractal basin boundaries.

1. Introduction

Belief propagation (BeP) is a relatively new and powerful method for inference and optimization problems. Inference problems arise in many fields of statistical physics, error-correcting codes and machine learning. Interestingly, some of the previously developed methods such as turbocodes or the transfer matrix approach in physics are in fact just variations of the same belief propagation technique. Inference problems deal typically with questions such as: Given a set of variables with statistical dependencies, what are their most probable states when only the states of a possibly small group of variables is known from data? As an example, think of the following system, taken from [1]: Holidays in an Asian country (A) increase the risk of tuberculosis (T), whereas smoking (S) can cause lung cancer (L) or bronchitis (B). Either (E) tuberculosis or lung cancer can be detected by an X-ray analysis (X), which, however, can not distinguish between both illnesses. Shortness of breath (Dyspnoea, D) can either be caused by bronchitis (B) or either (E) lung cancer or tuberculosis (see Fig. 1).

Such systems are called Bayesian networks and can be represented by graphs, where each variable corresponds to a node and the dependency between variables is denoted by interconnecting lines ("edges"). In our example, the dependencies between nodes are directed. Smoking increases the probability of lung cancer, but not vice-versa. In the context of Fig. 1, two probabilities are of main interest: The overall joint probability, \( p(x) \) that any event, i.e. illness, diagnosis etc. occurs, defined as

\[
p(x) = p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2)p(x_5|x_3)p(x_6|x_5)p(x_7|x_4)p(x_8|x_7)\]

...for example. This question deals with the marginal probabilities, which are calculated by summing over all possible states of a node’s parents:

\[
p(x_N) = \sum_{x_1} \sum_{x_2} \cdots \sum_{x_{N-1}} p(x_1, x_2, \ldots, x_N) \quad (1)
\]

From this definition, it is obvious that an exact mathematical computation is only practical for small graphs, since the number of terms in the sum grows exponentially with the number of nodes.

Pairwise Markov Random Fields (MRF) and Ising models: Consider the computer graphics task to infer some quantities of the underlying scene \( x_i \) (Fig. 2 a) from pixel image data \( y_i \). We further assume that there is a statistical relationship \( \phi(x_i, y_i) \) between \( x_i \) and \( y_i \) for each pixel \( i \). This function is often called evidence for \( x_i \), as it influences the probability for a scene site \( i \) to be of value \( x_i \), given that one...
observes \( y_i \). Additionally, in order to deduce anything at all from the image, there has to exist an underlying structure in the scene, expressed by a compatibility function \( \psi_{ij}(x_i, x_j) \). The compatibility function can be understood as a coupling between the scene quantities \( x_i \) (the word "coupling" is used here as it turns out that \( \psi_{ij} \) reduces to the physical coupling constant \( J \) for Ising systems).

Given these quantities and functions, we can write the overall joint-probability as

\[
p((x), (y)) = \frac{1}{Z} \prod_{ij} \psi_{ij}(x_i, x_j) \prod_i \phi(x_i, y_i) \tag{2}
\]

where \( Z \) is the normalization constant and the product runs over all connected \( x_i \)'s (for example, the nearest neighbours on a regular grid). In contrast to the Bayesian network discussed above, the MRF is undirected and pairwise, since the compatibility functions \( \psi_{ij} \) only depend on pairs of nodes \( i \) and \( j \). As with Bayesian networks, the computation of exact marginal probabilities is only possible for very small systems. The instructive part of MRF is how they relate to magnetic spin models. The energy of such a system is given by its spin-part Hamiltonian,

\[
E = -\sum_{i,j} J_{i,j} s_i s_j - \sum_i h_i s_i, \quad \text{where the first sum is taken over a site's nearest neighbours and } s_i \text{ is the spin of } i\text{'th site. Here, } J_{i,j}/T \text{ is identified essentially with } \log(\psi_{ij}). \text{ Its physical interpretation is that of a coupling strength between sites } i \text{ and } j. \log(\phi_i) \text{ is essentially identified with } h_i/T, \text{ following the physical interpretation of an external magnetic field. From statistical physics it is known that the states of the system in the canonical ensemble obey Boltzmann's law: The probability of finding the system in some spin configuration } \{s\} \text{ is } p(\{s\}) = \frac{\exp(\sum_{i,j} J_{i,j} s_i s_j + \sum_i h_i s_i)}{Z}. \text{ Comparing this to Eq. } (2), \text{ we realize that the MRF corresponds to a spin model at temperature } T. \text{ It also follows that the normalization constant } Z \text{ in } (2) \text{ can be interpreted as the system's partition function.}

**Belief propagation approach:** Because of their close relation to physics, we will focus in the following on MRF. Again, we are interested in calculating joint and marginal probabilities, as given by Eq. (2), but in an approximate manner only. In the following, we will identify \( \phi(x_i, y_i) \equiv \phi(x_i) \), since, for simplicity, we assume the observed nodes \( y_i \) to be fixed. The idea behind belief propagation is to approximate marginal probabilities for a site \( i \) by so-called beliefs \( b_i(x_i) \). The beliefs are calculated from messages \( m_{ji} \) sent to the \( i\)-th site from its neighbours \( j \):

\[
b_i(x_i) = k\phi(x_i) \prod_{j \in N(i)} m_{ji}(x_i) \tag{3}
\]

That is, we take the product over all messages coming in from the neighbouring sites \( N(i) \), multiply by the local evidence \( \phi_i \) and normalize by a constant \( k \) (all beliefs at a site have to sum to one). The messages \( m_{ji} \) are determined self-consistently by the iterative update rule

\[
m_{ji} \leftarrow \sum_{x_j} \phi_j(x_j) \psi_{ij}(x_i, x_j) \prod_{k \in N(i) \backslash j} m_{kj}(x_k). \tag{4}
\]

In other words, the new message site \( i \) is going to send to site \( j \) is determined by the messages that were previously sent to the \( i\)-th site from all its neighbours except \( j \), weighted by the "coupling strength" \( \psi_{ij} \) and the local evidence \( \phi_i \) (see Fig. 2 b). It is easy to prove that the beliefs \( b_i(x_i) \) converge to the exact marginal probabilities \( p_i(x_i) \) for singly connected graphs [1]. However, in contrast to the direct calculation of the marginals (1), computing time for belief propagation on singly connected graphs only grows linear with the number of connections between the nodes.

For loopy graphs, the situation is worse. On such graphs it is not guaranteed that belief propagation converges at all. For spin systems, it was shown that there exists a critical temperature \( T_C \) above which loopy belief propagation converges [2]. In this regime, the choice of initial messages is irrelevant and has only small influence on the required computing time. At lower temperatures, near and in the ferromagnetic phase, convergence, however, may not be achieved. It is important to note that if BeP converges, the beliefs correspond to stationary points of the Bethe free energy. This correspondence finally justifies the use of BeP for simulating thermodynamical systems such as the Ising model. A proof of this theorem can be found in [1].

### 2. BeP convergence and critical slowing down

The behaviour of BeP in dependence on the temperature is of great interest, e.g. for sequential superparamagnetic clustering (SSC, [3]). SSC is based on magnetic spin models and requires sweeps over large temperature ranges to come up with natural data clusters. The task can be accomplished by Monte Carlo simulations ([3]), but BeP is much faster (up to a factor of 20) - if it converges [4]. When clusters break apart, BeP slowing down is observed, where it can be shown that this temperature is exactly the critical temperature of the phase transition. To investigate this phenomenon, we will restrict ourselves to Ising systems, for they allow a message parametrization [2],

\[
tanh v_{ij} := m_{i,j}(-1) - m_{i,j}(-1), \tag{5}
\]

which simplifies the update rules significantly:

\[
tanh v_{ij}' = \frac{\tanh J_{ij} \sum_{k \in N(i)\backslash j} v_{kj})} {\sum_{k \in N(i)} v_{kj}}. \tag{6}
\]

Assuming uniform coupling \( J_{ij} = J \) and vanishing external field, two uniform fixed point solutions exist for Eq. (6), namely \( v_{ij} = v = 0 \) (zero magnetization, paramagnetic phase) and \( v_{ij} = v \neq 0 \) (non-vanishing magnetization, ferromagnetic phase), cf. Fig (3 a). For stability analysis, the Jacobi matrix can be calculated to be

\[
\frac{\partial v_{ji}'} {\partial v_{ji}} = \frac{1 - \tanh^2(h_j + \sum_{i \in N(j) / i} v_{ij})}{1 - \tanh^2(v_j')} \cdot \tanh J_{ij} \delta_{ji} \mathbf{1}_{N(j) / i}(k) \tag{7}
\]
It is straightforward to guess that a uniform initialization of the messages \( \nu_{ij} \) is an eigenstate of the regular grid with periodic boundary conditions: Let us assume that each site has \( q \) nearest neighbours. We can then write the tensor (7) as a \( q \cdot N \times q \cdot N \) matrix \( M \), where \( N \) is the total number of sites, by numerating the pairs \( ij \) as rows and the pairs \( (kl) \) as columns of \( M \). From (7) it is clear that each matrix row only contains \( q - 1 \) non-zero entries, since every node receives \( q - 1 \) messages only. With uniform message initializations, all \( q - 1 \) non-zero matrix entries per row take the value
\[
m := \frac{1 - \tanh^2((q - 1)\nu)}{1 - \tanh^2(J) \tanh^2((q - 1)\nu)} \tanh(J).
\]

Hence, the vector \((1, 1, \ldots, 1)\) is an eigenvector with eigenvalue \((q - 1)m\). In fact, this eigenvalue is the largest one.

**Proof:** Recall that a special form of Frobenius’ theorem states that if \( A_{nn} \geq 0 \) is irreducible, each of the following is true: 1) \( r = \rho(A) \in \sigma(A) \), where \( \rho \) is the spectral radius and \( \sigma(A) \) the spectrum of \( A \). 2) The unique vector \( p \) defined by \( Ap = rp \), \( p_i > 0 \) and \( \| p \|_1 = 1 \) is called the Perron vector. 3) There are no other positive eigenvectors. Since \( m > 0 \), the Jacobian (7) is clearly positive semi-definite. Irreducibility of a matrix, on the other hand, is equivalent to an underlying graph being strongly connected. A strongly connected graph is a graph where each pair of nodes can be connected by a sequence of directed edges. This is obviously the case in the Ising example, since the nodes and edges form a regular grid structure. Thus, the eigenvector \((1, 1, \ldots, 1)\) is the Perron vector and its eigenvalue is the largest one. This result valid for all grids having the same number of neighbours to each node, i.e. an infinite regular grid or a finite grid with periodic boundary conditions.

**Paramagnetic phase:** From the above theorem it follows immediately that the fixed point \( \nu = 0 \) is stable as long as \( J < \tanh^{-1}(1/(q - 1)) \) (Fig 3 b). The critical coupling/temperature is therefore \( J \approx 0.347 \) or \( T = 1/J \approx 2.89 \). This is in accordance to numerical simulations, where a typical critical slowing down can be observed in the BeP computing time (Fig. 3 c).

Moreover, it can be shown that BeP converges irrespective of the choice of initial conditions if \( |\sigma(A)| < 1 \) [2], where \( \sigma(A) \) is the spectral radius of the matrix
\[
A_{ijkl} = \tanh(J) \delta_{ij} \delta_{k0} / (1/k) \ .
\]
which is the message independent part of the Jacobi matrix (7). The condition guarantees BeP convergence for all (i.e. also non-uniform) initial conditions in the paramagnetic phase.

**Ferromagnetic phase:** In the ferromagnetic phase, despite having a stable, attractive fixed point (Fig. 3 b), convergence depends heavily on the choice of the initial messages \( \nu_{ij} \). Although physical solutions should break symmetry and switch to a uniform magnetization density \( \pm 1 \), BeP does not necessarily converge, according to (9).

Figure 3: a) Uniform fixed points \( \nu \) corresponding to the ferromagnetic and paramagnetic phases. b) Stability of the fixed points: \( T < T_C \) (\( \approx 2.89 \)): The ferromagnetic fixed point is stable (but not globally attractive, see text); \( T > T_C \): the paramagnetic fixed point becomes stable and globally attractive. c) A typical critical slowing down is observed at the 2nd order phase transition.

One of possibly many initial conditions for which BeP does not converge in the ferromagnetic phase is a situation as shown in Fig. 4 a). In the simulations, such a vortex was considered as a perturbation of the fixed points (6) on a \( 3 \times 3 \) regular grid with periodic boundary conditions.

In our simulations, the system was simulated for temperature sweeps below \( T_C \). The messages were initialized to either the stable ferromagnetic fixed point (without loss of generality, \( \nu > 0 \) in the following) or to the (in this temperature regime) unstable \( \nu = 0 \) fixed point. Before the simulation was run, a perturbation of variable strength directed towards the vortex was once applied (Fig. 4 b). BeP was then started, and the iterations required for convergence were counted. If more than 500 iterations were required, the system was assumed not to converge.

The criterion for convergence was that messages would differ by less than \( \varepsilon = 10^{-16} \) between successive updates. The same simulation procedure was then repeated for different temperatures and perturbation strengths \( J \). Since the temperature range considered belongs to the ferromagnetic phase, one would expect the perturbed system to converge to the \( \nu > 0 \) fixed point (Fig. 4 b).

This is indeed true when perturbations are applied to the ferromagnetic fixed point. A typical convergence plot for this situation is shown in Fig. 5 a), from which we see
that BeP converges for all perturbations except for the exact vortex initialization. In the latter case, the convergence dependence as a function of temperature seems to be inverted to the general case: In the ferromagnetic phase, the vortex initialization never converges. Upon temperature increase, the vortex, however, arrives at a critical point $T_C < T_C$, above which it always converges. Note that the vortex initialization also converges quickly at temperatures at which critical slowing down is observed in the general case. The previously mentioned "cut-off" at 500 iteration steps becomes reasonable when looking at the results in Fig. 5 a): First, a higher cut-off would only slightly narrow the peak around $T_C$. Second, the vortex initialization does not converge below $T_C$, even if we use significantly longer simulation runs.

A completely different behaviour is observed when we start with a perturbation of the unstable paramagnetic fixed point $v = 0$, as shown in Fig. 5 b). Another critical temperature seems to separate convergent from non-convergent regions at $T_C \approx 0.75$. Above this temperature, BeP converges, independently of the perturbation strength $\lambda$. Below $T_C$, convergent and non-convergent initializations can be found. The behaviour is non-trivial and depends strongly on the choice of the parameters $T$ and $\lambda$. Furthermore, as one would expect, for very small $\lambda$, BeP tends to converge more often. Apart from that, the choice of $\lambda$ does not influence the behaviour on a larger scale, in contrast to the temperature.

Below $T_C$, when zooming in the black square of Fig. 5 b), a self-similar structure is revealed. Even for high magnifications, a distinct boundary separating converging from non-converging regions is missing. It is an open question whether the nature of these findings lie in the theory of fractal basin boundaries.

3. Conclusion

Although loopy belief propagation on Ising systems has stable, attractive fixed points in both magnetic phases, it does generally only converge in the paramagnetic case. A BeP critical slowing down at $T_C \approx 2.89$ is observed at the phase transition, where both BeP fixed point solutions become unstable. In the ferromagnetic phase, vortex perturbations to the BeP fixed points reveal two entirely different behaviours: Perturbations of the stable ferromagnetic solution does not heavily influence BeP efficiency and convergence. With exact vortex initialization, BeP does not converge at low temperatures, whereas, interestingly, the vortex starts to converge to the paramagnetic phase at temperatures well below $T_C$. Perturbations of the unstable paramagnetic solution reveal a more complex situation. Convergence is usually achieved for $T > T_C \approx 0.75$ only. At lower temperatures, self-similar patterns are observed, the nature of which yet remains to be analyzed. An appealing explanation are fractal basin boundaries, separating the convergent from the non-convergent initial conditions.

References


Faster Spike Sorting with Belief Propagation

Thomas Ott, Albert Kern and Ruedi Stoop

Institute of Neuroinformatics, ETH/UNIZH Zurich, Switzerland
Email: tott@ini.phys.ethz.ch

Abstract—Clustering is a basic step common to most spike sorting methods. Sequential superparamagnetic clustering provides a powerful nonparametric clustering algorithm, especially if only little prior information is available. However, the traditional Monte Carlo-based implementation of this algorithm is relatively slow. We show that, by using the belief propagation method, superparamagnetic clustering is accelerated substantially (by a factor of 10-20). This makes it an attractive tool for spike sorting.

1. Introduction

Spike sorting is a fundamental step in the analysis of neural recordings. The spike sorting problem can be seen as a simpler analogue of the cocktail party problem in the field of auditory scene analysis. In both cases, we want to extract one or more signals from a mixture of different signals and noise. The goal of spike sorting is to detect spikes in extracellular voltage recordings and assign them to their different sources, i.e., the neurons.

Common spike sorting schemes are based on two ingredients [1]: (I) The description of the spike-underlying recorded waveforms by means of -usually high-dimensional- vectors, and (II) a clustering algorithm that identifies groups of similar waveforms. Spikes in each group are assumed to originate from the same neuron. This is a reasonable assumption if the underlying processes are stationary and the number of spike overlaps is small. For the clustering algorithm, two often conflicting properties are desirable. Firstly, the algorithm should be as unbiased as possible, i.e., it should not rely on prior information about the number or the shape of the clusters and it should also work for high dimensions. Secondly, it should be reasonably fast.

In this contribution, we apply the sequential superparamagnetic clustering algorithm (SSC) [2] to a) an artificially created and to b) a real spike sorting problem. SSC satisfies the requirement of an unbiased algorithm, but it is rather slow. We discuss how SSC is considerably accelerated by means of belief propagation (BeP) [3]. Finally, the time performance of BeP-mediated SSC is compared to that of Monte Carlo-based SSC.

2. The Basic Stages of Spike Sorting

Extracellular recordings by a single electrode provide a waveform train made up of the activity of an unknown number of neurons and of background noise. In a first step, potential spike shapes are identified by a thresholding procedure: If the recorded voltage crosses a threshold, a sample vector (wave trace) containing the voltage values over a certain time window length is selected. In our example (Fig. 1), the sample vector is of dimension 48. The choice of the threshold is delicate. For too small thresholds, noise artifacts may be detected and noise fluctuations can lead to a bad alignment of the peaks. Although for the latter can be corrected by a subsequent alignment, a fixed window size may still cause problems. If the threshold is too large, some spikes might be missed. In practice, one tries to use relatively large thresholds that still catch all spikes. In a next stage, the spike traces need to be characterised in the form of feature vectors which then allow for a comparison and cluster identification. For this purpose, different strategies have been introduced. A simple characterisation of the spikes can be achieved by a small number of characteristic features such as spike height or spike width. Sometimes, the whole sample vector is used for a comparison by means of the Euclidian distance. More frequently, the dimension is reduced by a principal component analysis (PCA) that can be considered an automatic feature selection. This sometimes allows for a straightforward detection of clusters by eye on the basis of the first two principal components. Alternatively, the wave trace may be characterised by Fourier- or Wavelet coefficients [4], or by coefficients of any other transform that works out suitable.

1Data from a randomly chosen set of recordings from macaque parietal cortex provided by H. Scherberger and collaborators.

Figure 1: A set of 241 wave traces from extracellular recordings of a single electrode. The time resolution is 30 kilo-samples/s, the amplitude is measured in μV.
characteristics. In short, the clustering preprocessing phase provides us with feature vectors $v_i$ whose distance is given by an appropriate measure. In the final stage, we can thus perform the actual clustering. Although we should keep in mind that the preprocessing is of equal importance, we entirely focus on this last stage in the following. The prospects of a successful spike sorting vanish if a cluster spreads over the whole feature space or if different clusters largely overlap. In such cases, the preprocessing must be revised. Generally, we are confronted with the difficulty that the number of clusters, i.e., the number of neurons recorded, is unknown. Another difficulty is that the shape of the clusters can be complicated and the densities of the clusters may differ considerably. Practically, the lack of prior information suggests the usage of flexible non-parametric clustering algorithms. From a conceptual point of view, it seems natural to think of nonparametric clustering as a self-organisation process; e.g., in a network of interacting elements, clusters can emerge naturally as global characteristics from local interactions. Several clustering algorithms based on self-organisation processes have been introduced during the last years (e.g., [5]). Superparamagnetic clustering (SC) [6] is an algorithm which has attracted particular attention in many fields of application. The algorithm is based on the pattern formation process a Potts spin system undergoes within a certain temperature range. SC has been substantially extended to sequential superparamagnetic clustering (SSC) [2] in order to make it more robust in the case of highly inhomogeneous clusters. Additionally, SSC provides us with a measure for the automatic selection of the most natural clusters.

3. Superparamagnetic Clustering with Belief Propagation

SC maps a clustering problem onto the setting of a Potts spin system. Each feature vector $v_i$ is represented by a spin variable $s_i$ that can take on a value in $\{1, \ldots, q\}$. In the following, we set $q = 2$ and re-denote the two possible spin values by $\{-1, 1\}$ which constitutes the more familiar case of Ising spins. Each spin is coupled to its $k$ nearest neighbours and the coupling strengths are chosen according to

$$J_{ij} = \frac{1}{k} \exp \left( -\frac{d_{ij}^2}{2a^2} \right),$$

where $a$ is the average distance of coupled neighbours and $d_{ij}$ is the distance between $v_i$ and $v_j$.

The probability of a spin configuration $s$ is given by the Boltzmann distribution

$$p(s) = \frac{1}{Z(T)} e^{-H(s)/T},$$

with the Hamiltonian $H(s) = -\sum_{i,j} J_{ij} s_i s_j$. Depending on the temperature $T$, two spins in an inhomogeneous Potts spin system can either be correlated or uncorrelated. The change from the correlated to the uncorrelated state expresses a collective phenomenon and occurs simultaneously for whole groups of spins at (finite-type) phase transitions. Increasing $T$ from zero leads the system through different phases. For small $0 \leq T < T_{\text{ferro}}$, the system is in the ferromagnetic phase. In other words, all spins form one single cluster. For $T_{\text{ferro}} < T < T_{\text{max}}$, the system is in the superparamagnetic phase, where groups of correlated spins form local clusters. Upon the increase of $T$, the local clusters break up into ever more and smaller clusters, in which only the most strongly coupled spins remain correlated (clustering hierarchy). At $T = T_{\text{max}}$, the system enters the paramagnetic phase and the spins are no longer correlated.

From the sketched picture it is apparent that the basic quantities to be computed for clustering are the pairwise correlations

$$c_{ij} = \sum_t p(s) \delta_{s_i s_j}.$$  \hspace{1cm} (3)

For the actual clustering algorithm, clusters are defined as sets of pairwise aligned spins. Two spins are called aligned if $c_{ij}$ exceeds the threshold $\Theta = 0.7$ (the exact choice of $\Theta$ is uncritical [6]).

Traditionally, $c_{ij}$ was evaluated by means of Monte Carlo methods. Recently, a novel method, belief propagation (BeP), has come up which generally allows for a fast calculation or approximation of marginal probabilities [3]. We show that accordingly BeP can be employed to calculate $c_{ij}$. However, the convergence properties of BeP are non-trivial and care has to be applied in order to assure that BeP converges. BeP is based on the exchange of messages $m_{j\rightarrow i}(s_j)$ and $m_{i\rightarrow j}(s_i)$ between pairs of coupled spins $i$ and $j$, containing a recommendation about in what state the other spin should be. Given the set of messages at time $t$, $\{m_{j\rightarrow i}(s_j)\}$, the outgoing messages at time $t + 1$ are determined by the weighted products of incoming messages

$$m_{i\rightarrow j}^{t+1}(s_j) = k_1 \sum_{s_i} e^{2J_{ij}s_is_j} \prod_{keN(i)\setminus j} m_{k\rightarrow i}^t(s_i),$$

where $k_1$ is a normalisation constant assuring that $\sum_{s_i} m_{i\rightarrow j}^{t+1}(s_j) = 1$ and $N(i)\setminus j$ denotes the set of all neighbouring sites of $i$ without $j$. Once BeP has reached a fixed point $\{m_{i\rightarrow j}^\infty(s_j)\}$, $c_{ij}$ can be approximated by means of the pairwise beliefs $b_{ij}(s_i, s_j)$, where $c_{ij} = b_{ij}(1, 1) + b_{ij}(-1, -1)$ and

$$b_{ij}(s_i, s_j) = k_2 e^{2J_{ij}s_is_j} \prod_{keN(i)\setminus j} m_{k\rightarrow i}^\infty(s_i) \prod_{leN(j)\setminus i} m_{l\rightarrow j}^\infty(s_j).$$

$k_2$ again is a normalisation constant.

The set of the pairwise beliefs $b_{ij}$ and of the single beliefs $b_i = \sum_j b_{ij}$ obtained from a stable fixed point of (4) corresponds to a local minimum of the Bethe free energy functional [3, 7]. For spin systems with a loopy grid structure, the Bethe free energy is an approximation of the actual free energy.
energy functional which is minimised by the Boltzmann distribution. The Bethe free energy ansatz assumes that the free energy is a functional of the marginal probabilities \( p_j(s_i, s_j) = \sum_{j', k} p(s) \) and \( p_j(s_i) = \sum_j p_j(s_i, s_j) \). Generally, this is not quite correct, but nevertheless, a minimum of the Bethe energy often gives us a fairly good approximation for the exact marginals. In general, it is however impossible to get exact values by means of BeP. While for clustering applications this approximative character of BeP is acceptable and works well, the problem of convergence is more severe: BeP convergence is only guaranteed in the paramagnetic phase, independent of the message initialisation. In the ferro- or superparamagnetic phase, BeP can get trapped in cycles [8] or may possibly settle on a chaotic attractor. Fortunately, we found a set of message initialisations for which convergence of BeP seems very reliable (This statement can be confirmed by analytical considerations [8]). The messages \( m_{i\rightarrow j} \) have to be initialised uniformly according to

\[
m^0_{i\rightarrow j}(1) = y \text{ and } m^0_{i\rightarrow j}(-1) = 1 - y,
\]

with the same \( 0.5 < y < 1 \) for all connections \( i \rightarrow j \).

4. Sequential Superparamagnetic Clustering SSC

Usually, we are interested in detecting only the most natural or obvious clusters. Clusters that are due to some random local grouping in the data are considered as noise. Intuitively, we have to sweep through the superparamagnetic phase by increasing \( T \) and identify the most natural clusters as the ones that persist over a large temperature range \( T_{cl} \) (coarse-grained structure). These clusters are well reflected in the BeP-approximation, which is the reason why BeP works well for SC/SSC. However, the clustering solution is generally compromised if the clusters to be detected are characterized by largely differing densities [2]. By extracting the reliable clusters and reclustering the remainder, SSC iteratively re-generates reliable coarse-grained structures on finer levels. The desired level of refinement is the main parameter set by the user \( (s_0 \in [0,1]) \). It corresponds to a minimal required length of the cluster existence interval \( T_{cl} \) in comparison to \( T_{max} \). Additionally, \( T_{fer} \) of a clustered set is a good measure for its intrinsic cluster coherence (its “clusterness”). A detailed description of SSC can be found in [2]. For relatively simple data sets with clear cluster structures, the clusters detected by SSC agree with the coarse grain structures obtained by superparamagnetic clustering SC.

5. Performance

Sorting Performance: The advantages of superparamagnetic clustering (or the further developed SSC) if compared to the often used K-means algorithm (or similar algorithms) are obvious: We do not need to know the number of clusters and there is no shape bias. Only if the number of clusters is known and their shape is spheric, then it might be advantageous to use K-means. Unfortunately, a direct comparison of the performance of different algorithms based on real-world data recordings is difficult. Either the correct sorting is unknown and a standard solution to compare with is missing, or the sorting is rather simple and all algorithms perform well. The advantages and superiority of SC and particularly SSC, however, has been demonstrated in connection with other applications (see [2] and references therein). Here, we do not provide a systematic discussion, but only give an impression by means of two examples.

In Fig.2, SSC has been applied to an artificially created data set with 3 clearly distinguishable basic waveforms\(^4\). The 3 classes are easily recognised by SSC without specifying any parameters (Fig.2b - 2d).

In Fig. 3, we show the clustering result obtained by SSC for the set shown in Fig. 1. The clustering was performed on the first two principal components PCA1 and PCA2 only. A fast coarse-grained clustering with a large \( s_0 = 0.5 \) yields 2 clusters (Fig. 3a). Cluster 1 (Fig. 3b/3c) is the purer cluster as it is extracted first. Cluster 2 (Fig. 3d) is the residual cluster. It still contains some clear outliers (Fig. 3d). Fortunately, SSC provides information about the reduced purity of cluster 2 in the form of \( T_{fer} \). \( T_{fer} \) for cluster 2 is 10 smaller than \( T_{fer} \) for cluster 1 (Fig. 3a). Consequently, by using a smaller \( s_0 \), cluster 2 will be further decomposed into a core cluster and a residual cluster (cluster halo) with inherent marginal intrinsic coherence (small \( T_{fer} \); results not shown).

\(^4\)This data set testspikes.mat is provided by R.Q.Quiroga and can be downloaded under www.vis.caltech.edu/~rodr
the Swendsen-Wang algorithm at least about 200 Monte Carlo steps are necessary to obtain stable results, whereas BeP provides satisfying solutions within as few as 10 to 20 iterations. This seems generically true, although the effective speed of convergence differs along the $T$-axis and, in particular, critical slowing down is observed at points of cluster break-up [8]. As a main result of this contribution, we state that BeP leads to a substantial speed up of clustering by at least a factor of 10. With an optimal implementation we expect a speed-up factor of 20.

6. Summary and Outlook

In this contribution, we worked out the prospect of sequential superparamagnetic clustering as an attractive and unbiased tool for spike sorting with clear advantages over standard clustering algorithm such as K-means. We showed that the algorithm is accelerated substantially by means of belief propagation which renders also its time performance competitive. In our future work, a whole spike sorting method has to be elaborated taking into account the clustering preprocessing stage. More systematic comparisons will certainly allow to substantiate the advantages of the proposed clustering approach. We are confident that our approach possesses a great potential for a wide-spread usage.

Acknowledgements

We would like to thank H. Scherberger, M.-C. Fluet and M. Baumann from the Institute of Neuroinformatics for providing samples of neural recordings. This research has been supported by a ZNZ (Neuroscience Center Zurich) grant.

References


**Bifurcation structure of coexisting two duck solutions and their breakdown into chaos**

Munehisa Sekikawa$^{1,2,*}$, Naohiko Inaba$^3$ and Kazuyuki Aihara$^{2,1}$

$^1$ Aihara Complexity Modelling Project, ERATO, JST; 3-23-5-2F Uehara, Shibuya-ku 151-0064, Japan
$^2$ Institute of Industrial Science, The University of Tokyo; 4-6-1 Komaba, Meguro-ku 153-8505, Japan
$^3$ Faculty of Engineering, Utsunomiya University; 7-1-2 Yoto, Utsunomiya-shi 321-8585, Japan
$^*$E-mail: muse@aihara.jst.go.jp

**Abstract**—In this paper, we discuss the generation of coexisting two duck solutions and their breakdown into chaos in a singularly perturbed van der Pol oscillator driven by extremely small periodic force. Duck solution breakdown into chaos via period-doubling bifurcations cascade is also observed.

1. Introduction

The amplitude of a limit cycle sometimes expands rapidly when a system parameter is varied slightly in a singularly perturbed system with a small parameter $\varepsilon$. This rapid change from a small amplitude oscillation to a large amplitude oscillation is known as the canard phenomenon or the duck solution, and the phenomenon has been investigated in various fields of science [1]–[8]. The duck solution was originally discovered by Benoît et al. for the following singularly perturbed van der Pol (vdp) oscillator [1]:

$$
\begin{cases}
\varepsilon \dot{x} = y + x(1 - x^2) \\
\dot{y} = -x + B_0.
\end{cases} 
$$

According to the results of the nonstandard analysis, it is clarified that the amplitude of the limit cycle is changed by the order of 1 when the parameter $B_0$ is varied by the order of $\exp(-1/\varepsilon)$ [2, 3]. $\exp(-1/\varepsilon)_{\varepsilon \to 0} \approx 4.5 \times 10^{-5}$ and $\exp(-1/\varepsilon)_{\varepsilon = 0.01} \approx 3.7 \times 10^{-41}$. The amplitude is extremely sensitive to $B_0$.

What influence does the perturbation of $B_0$ exert on the duck solution? This problem seems to be an attractive hot topic [4, 7], since the solution is hypersensitive to the fluctuation of the $B_0$ and noise exists in the actual system.

In this paper, we analyse the forced vdp equation in the following form:

$$
\begin{cases}
\varepsilon \dot{x} = y + x(1 - x^2) \\
\dot{y} = -x + B_0 + B \sin(\omega t),
\end{cases} 
$$

where the value of $B_0$ is chosen such that the duck solution is generated when $B = 0$, and the value of $B$ is chosen to be on the order of $\exp(-1/\varepsilon)$. Ultraharmonic, subharmonic and fractional harmonic (ultra-subharmonic) entrainments are observed if $B$ is small. By the analysis of Arnold’s tongue, we find a very interesting bifurcation structure for these entrainments. These entrainments are generated by saddle-node bifurcations. In each entrainment, the shape of the attractor is similar to that of a duck. In succession, another saddle-node bifurcation occurs, and another duck solution is generated. Namely, two duck solutions coexist in Arnold’s tongue. The duck solution breakdown into chaos via period-doubling bifurcations cascade is also observed in each entrainment.

2. Results

Since the perturbation is assumed to be periodic, the Poincaré map $T_\lambda$ is defined as follows:

$$
T_\lambda : \mathbb{R}^2 \to \mathbb{R}^2 \\
u \mapsto T_\lambda(u) = \varphi(2\pi/\omega, u, \lambda),
$$

where $\lambda$ is one of the parameters of $(\varepsilon, B_0, B, \omega)$, and $\varphi(t) = (x(t), y(t))^T$ is a solution. A synchronous solution of the continuous dynamic system (2) corresponds to a periodic point of $T_\lambda$. Let us denote an $m$-periodic point by $1d^m$ or $1d^m$. $d$ and $i$ correspond to the periodic points which have even and odd number of the characteristic multiplier on the real axis ($-\infty, -1$), respectively, and $k$ indicates the number of the characteristic multiplier outside the unit circle in the complex plane [9]. $1d^m$ is a stable $m$-periodic point, $1d^m$ and $1d^m$ are a saddle, and $3d^m$ is a completely unstable periodic point.

Figure 1: Duck solution observed in Eq. (2) with $\varepsilon = 0.1$, $B_0 = 0.5694$ and $B = 0.0$. 

1031
The following bifurcations are observed in Eq. (2): A **saddle-node bifurcation** \((G)\) occurs when one of the eigenvalues of an \(m\)-periodic point is 1. If the parameters are varied in order to pass through the bifurcation set, a saddle and a stable periodic point merge together and these periodic points vanish:

\[
\phi \xleftarrow{G} 1d^m + \omega d^m,
\]

where the symbol \(\xleftarrow{G}\) indicates the relation before and after the bifurcation and \(\phi\) denotes the extinction of the periodic points. Hereafter, this bifurcation parameter set is denoted as \(G\).

A **Period-doubling bifurcation** \((I)\) occurs when one of the eigenvalues of an \(m\)-periodic point is \(-1\). A stable \(m\)-periodic point becomes a saddle, and a stable \(2m\)-periodic point is generated near this saddle:

\[
\omega d^m \xleftarrow{I} 1d^m + \omega d^{2m}.
\]

A period-doubling bifurcation occurs successively. It is well known that one of the routes to chaos is the period-doubling bifurcations cascade. Hereafter, this bifurcation parameter set is denoted as \(I\). The bifurcation values of the system parameters are obtained by Kawakami’s procedure [9].

In the following discussion, \(\varepsilon\) is fixed at 0.1 and \(B_0\) is fixed at 0.5694. The duck solution appears at these values of the parameters. It is illustrated in Fig. 1. Figure 2 shows the 2-parameter bifurcation diagram. The abscissa denotes the angular frequency of the periodic perturbation \(\omega\) and the ordinate denotes its amplitude \(B\). Bifurcation sets drawn in red and in green denote the saddle-node bifurcations \((G)\), and bifurcation sets drawn in blue denotes the period-doubling bifurcations \((I)\).

In Fig. 2, 1/1 denotes the fundamental harmonic entrainment. \(n/1, 1/m\) and \(n/m\) denote the \(n\)th ultraharmonic, 1/m subharmonic and \(n/m\) fractional (ultra-subharmonic) entrainments, respectively. The maximum Lyapunov exponent is calculated by Shimada-Nagashima algorithm [10], and chaos-generating regions are shaded in this figure. According to this figure, we can see that chaos is observed for \(B \geq \exp(-1/\varepsilon))_{\varepsilon=0.1} \approx 4.5 \times 10^{-5}\).

Figure 3 shows magnified pictures of the ultraharmonic entrainments, subharmonic entrainments and fractional harmonic entrainments. We find that two types of bifurcation structure exist. The schematic picture of the first type is illustrated in Fig. 4. Figures 3 (e), (i), (j) and (k) belong to this type. In this schematic figure, a duck solution and a duck-shaped saddle, which is a saddle attractor with a duck-like shape on the phase plane, are generated by a saddle-node bifurcation \((G_1\) in red). By increasing the parameter \(B\), the other saddle-node bifurcation \((G_2\) in green) occurs inside the entrainment and then the other duck solution and duck-shaped saddle are generated. These two pairs of the duck solution and the duck-shaped saddle exist, where \(B\) is in the order of \(\exp(-1/\varepsilon)\). Figure 5 shows the coexisting duck solutions observed in the fundamental harmonic entrainment. According to Fig. 4, the coexisting two duck solutions and two duck-shaped saddles constitute an identical periodic point manifold. One duck solution which is generated by \(G_1\) remains stable against the perturbation in the range where \(B\) is on the order of \(\exp(-1/\varepsilon)\). On the other hand, the other duck solution which is generated by \(G_2\) becomes unstable with increasing in parameter \(B\) and bifurcates to chaos via period-doubling bifurcations. Figure 6 shows attractors observed in the fundamental harmonic entrainment. The period-doubling bifurcations cascade and chaos are observed in the range where \(B\) is on the order of \(\exp(-1/\varepsilon)\).

A schematic picture of the second-type entrainment is illustrated in Fig. 7. Figures 3 (a), (b), (c), (d), (f), (g) and (h) belong to this type. In this type of bifurcation diagram,
Figure 3: Magnified pictures: (a) 3rd ultraharmonic entrainment; (b) $8/3$ fractional harmonic entrainment; (c) $5/2$ fractional harmonic entrainment; (d) $7/3$ fractional harmonic entrainment; (e) 2nd ultraharmonic entrainment; (f) $5/3$ fractional harmonic entrainment; (g) $3/2$ fractional harmonic entrainment; (h) $4/3$ fractional harmonic entrainment; (i) fundamental harmonic entrainment; (j) $1/2$ subharmonic entrainment; (k) $1/3$ subharmonic entrainment.

Figure 4: Schematic picture of periodic point manifold of first type.

Figure 5: Coexisting two duck solutions observed in fundamental harmonic entrainment with $K = 0.1, B_0 = 0.5694, \omega = 1.19, B = 5 \times 10^{-3}$ (at point A in Fig. 3(i)).
Figure 6: Attractors observed in fundamental harmonic entrainment with $\epsilon = 0.1$, $B_0 = 0.5694$, $\omega = 1.19$: (a) periodic attractor with period 1 with $B = 5.5 \times 10^{-5}$; (b) periodic attractor with period 2 with $B = 6.0 \times 10^{-5}$; (c) periodic attractor with period 4 with $B = 6.2 \times 10^{-5}$; (d) chaotic attractor with $B = 6.5 \times 10^{-5}$.

although a duck solution and a duck-shaped saddle are generated by $G_1$ and $G_2$, the duck solution is collapsed immediately by the increase in parameter $B$, and chaos is generated by the period-doubling bifurcations cascade. Coexisting two duck solutions are also observed, but the coexisting regions are relatively narrow. As shown in Figs. 3 (b), (c) and (d), we could not observe the second saddle-node bifurcation in the range where $B$ is on the order of $\exp(-1/\epsilon)$.

3. Conclusion

In this study, we investigated the response of duck solution driven by weak periodic force. The bifurcation structure of each entrainment is revealed in detail. The coexisting two duck solutions is observed under an extremely small periodic perturbation. These duck solutions collapse and chaos via period-doubling bifurcations cascade are also observed in the regions where the amplitude of the periodic force is extremely small.

Acknowledgments

We would like to thank Prof. H. Kawakami and Prof. T. Yoshinaga for valuable comments on this research.

References

Analysis of Co-existence Phenomena of Superstable Periodic orbit and Chaos in a Nonautonomous Piecewise Constant Circuit

Yusuke Matsuoka† and Toshimichi Saito†

†EECE Dept., HOSEI Univ., 3-7-2 Kajino-cho, Koganei-shi, Tokyo, 184-8584 Japan
Email: {matsuoka@nonlinear., tsaito@ } k.hosei.ac.jp

Abstract—This paper studies complicated superstable phenomena in a nonautonomous piecewise constant circuit. The circuit equation has piecewise constant vector field and piecewise linear trajectory: it is well suited for theoretical analysis. The vector field causes rich bifurcation phenomena including co-existence of chaos and superstable periodic orbits. As a parameter varies, we can see rich bifurcation phenomena: chaos is changed into superstable periodic orbits and so on. Using the embedded return map, we analyze bifurcation phenomena precisely.

1. Introduction

This paper studies a non-autonomous piecewise constant (PWC) circuit having one impulsive switch. The switching causes rich bifurcation phenomena. The circuit equation is transformed into Equation (3).

\[ \begin{align*}
\dot{x} &= -\text{sgn}(x - ay) \\
\dot{y} &= -\text{sgn}(x) \\
(x(t^+), y(t^+)) &= (-1, y(t)) \quad \text{at } \tau = t.
\end{align*} \]

Figure 1 shows a non-autonomous PWL circuit model.

1. Introduction

This paper studies a non-autonomous piecewise constant (PWC) circuit having one impulsive switch. The switching causes rich bifurcation phenomena in various systems: simplified model of DC/DC converters, simple pulse-coupled neuron models and flat-topped tent maps [8]-[10]. Note that [11] firstly present co-existence of SSPOs and chaos, however, related bifurcation phenomena have not been analyzed sufficiently. Analysis of this paper may be fundamental result for studying pulse-coupled neural networks [4], [6], [9] and so on.

2. Circuit model and typical phenomena

Figure 1 shows a non-autonomous PWL circuit model. The circuit consists of two VCCSs, two capacitors and one time-dependent impulsive switch \( S \) where \( r_1 \) and \( r_2 \) denote parasitic resistors and \( T \) is a period of clock signal. The VCCSs have signum characteristics:

\[ \begin{align*}
\{ i_1 &= I_1 \text{sgn}(v_1) \\
\dot{v}_2 &= I_2 \text{sgn}(v_1 - v_2)
\end{align*} \]

where \( I_1 < 0 \) and \( I_2 < 0 \). At every period \( t = nT \), \( S \) is closed and capacitor voltage \( v_1(t) \) is reset to the dc voltage \( E < 0 \). For simplicity, we assume that the switching is instantaneous without delay and continuity property of \( v_2(t) \) is held. We consider the case where \( r_1 \) and \( r_2 \) are large enough and open them for simplicity. In this case the circuit dynamics is described by

\[ \begin{align*}
C_1 \frac{dv_1}{dt} &= I_2 \text{sgn}(v_1 - v_2) \\
C_2 \frac{dv_2}{dt} &= I_1 \text{sgn}(v_1)
\end{align*} \]

(1)

where \( n \) is a non-negative integer. Using the dimensionless variables and parameters, Equation. (2) is transformed into Eq. (3).

\[ \begin{align*}
\dot{x} &= -\text{sgn}(x - ay) \\
\dot{y} &= -\text{sgn}(x) \\
(x(\tau^+), y(\tau^+)) &= (-1, y(\tau)) \quad \text{at } \tau = nd.
\end{align*} \]
For simplicity, we focus on the parameter range : 
\[ a < d < a + 1. \]  
(6)

For simplicity, let us consider the parameter range :
\[ a < d < a + 1. \]  
(6)

This system has two positive real parameters \( a \) and \( d \). Since Eq. (3) expresses PWC vector field, the trajectory is PWL. For simplicity let us consider the parameter range :
\[ a > 1, \quad d > 0. \]  
(4)

Typical circuit dynamics is shown in Fig. 2 (a) and (b). In this case the trajectories draw stable rect-spiral when \( S \) is open, and jump to the base line \( x = -1 \) instantaneously at the moment when \( S \) is closed. Repeating in this manner, the circuit exhibits various interesting phenomena as shown in Fig. 2 (c) to (f). It should be noted that the trajectory can reach an equilibrium point within some finite time. 

In order to analyze the phenomena, we define one-dimensional return map. Let \( L_D = \{(x, y, \tau) \mid x = -1 \text{ and } \tau = nd \} \) and let a point on \( L_D \) be represented by its \( y \) coordinate as shown in Fig. 3. We assume that the trajectory starts from \( L_D \) after the switching at \( \tau = 0 \). The \( n \)-th switching generates at \( \tau = nd \). Since \( y_n \in L_D \) and \( y_{n+1} \in L_D \) are satisfied, we can derive the one-dimensional return map \( F \) from \( L_D \) to itself:
\[ F : L_D \rightarrow L_D, \quad y_{n+1} = F(y_n), \quad y_n \equiv y(nd). \]  
(5)

For simplicity, we focus on the parameter range :
\[ a < d < a + 1. \]  
(6)

Figure 2: Basic dynamics and typical attractors for \( a = 2.2 \). (a) Behavior of the trajectory in time-domain. CLK is a clock signal. (b) Behavior of the trajectory in phase plane. (c) Complicated SSPOs for \( d = 2.5 \). (d) Chaos for \( d = 2.9 \). (e) and (e’) Co-existence of chaos and complicated SSPOs for \( d = 3.1 \), respectively. (f) Basic SSPO for \( d = 3.3 \).

Figure 3: Definition of the return map in phase space.

Figure 4: Typical return maps for \( a = 2.2 \). (c) Complicated SSPOs for \( d = 2.5 \). (d) Chaos for \( d = 2.9 \). (e) Co-existence of chaos and complicated SSPOs for \( d = 3.1 \). (f) Basic SSPO for \( d = 3.3 \). SFP denotes superstable fixed point. (c) to (f) correspond to Fig. 2 (c) to (f), respectively.

[7] Such reachability occurs by signum VCCSs and rich bifurcations appear as shown in Section 4.

3. Return maps with one flat segment

In order to analyze the phenomena, we define one-dimensional return map. Let \( L_D = \{(x, y, \tau) \mid x = -1 \text{ and } \tau = nd \} \) and let a point on \( L_D \) be represented by its \( y \) coordinate as shown in Fig. 3. We assume that the trajectory starts from \( L_D \) after the switching at \( \tau = 0 \). The \( n \)-th switching generates at \( \tau = nd \). Since \( y_n \in L_D \) and \( y_{n+1} \in L_D \) are satisfied, we can derive the one-dimensional return map \( F \) from \( L_D \) to itself:
\[ F : L_D \rightarrow L_D, \quad y_{n+1} = F(y_n), \quad y_n \equiv y(nd). \]  
(5)

For simplicity, we focus on the parameter range :
\[ a < d < a + 1. \]  
(6)
where $y_f$ is superstable periodic point, $F^k$ denotes the $k$-fold composition of $F$ and $k$ is a positive integer. Note that SSPO is superstable for initial state, but sensitive for parameters as shown afterward. When $k = 1$, $l$ does not mean and the map has superstable fixed point (ab. SFP) corresponding a basic SSPO (with period one) as shown in Fig. 2 (f) and Fig. 4 (f). When $k \geq 2$, $1 \leq l < k$ is satisfied and the map does not have SFP. The orbits hit the flat segment after hitting many branches. These orbits are said to be complicated SSPOs (see Fig. 2 (c) and 4 (c)).

4. Analysis of bifurcation phenomena

Fig. 5 shows bifurcation diagrams as the clock period $d$ varies and $a = 2.2$ is fixed, where $d$ increases in Fig. 5

Figure 5: Bifurcation diagrams for $a = 2.2$. (a) Bifurcation diagram as $d$ increases from $d = 2.2$. (b) Bifurcation diagram as $d$ decreases from $d = 3.2$. (c) is an enlargement of a dotted box in (b). B1 to B4 mean the borders of bifurcation, respectively.

Since the trajectory is PWL, the return map is PWL and is given explicitly. The map can be described as shown in [7]:

\[
F(y) = \begin{cases} 
  y - (d - 2) & \text{for } y \geq D_2 \\
  (-1)^n(\beta_{n+1}(y + 1) + 1 - d) & \text{for } D_{n+1} > y \geq D_{n+2} \\
  (-1)^n(\kappa_n(y - 1) + 1 + d) & \text{for } D_{n+1} > y \geq D_{n} \\
  y + d & \text{for } D_1 > y
\end{cases}
\]

where $\alpha \equiv \frac{\omega^{1}}{\pi} \approx 1 + 2 \alpha + \cdots + 2 \alpha^n$, $\beta_{n+1} \equiv 2 \alpha^n - 1$, $\beta_{n+2} = -\frac{1}{\alpha(n-d)} - 1$, $\beta_{n+1} = \frac{1}{\alpha(n-d)} + 1$, $D_{n+1} = \frac{1}{\alpha(n-d)} - 1$, $D_{n+2} = \frac{1}{\alpha(n-d)} + 1$ and $n$ is a positive integer.

Typical return maps are illustrated in Fig. 4. The map has one flat segment for $D_{n+1} < y_n < D_{n+2}$ and infinite extrema. This segment causes various complicated SSPOs presented in Fig. 4 (c) and (e). Here we indicate the formal definition of SSPOs. If Eq. (8) is satisfied, the system exhibits SSPOs with period $k$.

\[
F^k(y_f) = y_f, \quad F^l(y_f) \neq y_f, \quad \frac{d}{dy}F^k(y_f) = 0, \quad (8)
\]
(a) and d decreases in Fig. 5 (b) and (c), respectively. We can see a variety of bifurcation phenomena on the same borders. We then focus on the four typical bifurcations and consider their mechanism precisely.

On border B1 in Fig. 5, we can see that chaotic orbit is changed into complicated SSPOs. The return maps at the left and right of B1 are shown in Fig. 6 (a) and (b), respectively. Fig. 6 (a) has one invariant interval $J_1$ such that $F(J_1) \subseteq J_1$ and $|F(y_p) - |y_p|| > 1$ on $J_1$ where $J_1 \equiv [F(D_1), F(D_2)]$. In this case the map exhibits chaotic orbit on $J_1$ [2]. As $d$ increases and passes through B1, $J_1$ is to include the flat segment as shown in Fig. 6 (b). The orbits are stabilized if the orbits hit the flat segment, afterward the orbits become complicated SSPOs. The return maps at the near left and right of B2 are shown in Fig. 6 (c) and (d), respectively. Fig. 6 (c) exhibits complicated SSPOs. As $d$ increases and passes through B1, $J_1$ is to include the flat segment, afterward the orbits become complicated SSPOs. The border is given by $B_1 = [a, d] D_{\infty} = F(D_2))$. On this border, $d \approx 2.375$ for $a = 2.2$.

Near border B2 in Fig. 5, we can see that complicated SSPOs are changed into chaotic orbit. The return maps at the left and right of B2 are shown in Fig. 6 (c) and (d), respectively. Fig. 6 (c) exhibits complicated SSPOs. As $d$ increases and passes through B2, the map is to have one invariant interval $J_2$ where $J_2 \equiv [\bar{F}(D_2), F(D_3)]$. In this case the map exhibits chaotic orbit on $J_2$. The border is given by $B_2 = [a, d] D_{\infty} = F(D_3))$. On this border, $d \approx 2.806$ for $a = 2.2$.

On border B3 in Fig. 5, the map at the left of B3 is shown in Fig. 4 (e). This map has only invariant interval $J_3$. As $d$ increases and passes through B3, the map has two invariant intervals $J_2$ and $J_3$ as illustrated in Fig. 6 (f) where $J_3 \equiv [F(D_3), F(D_4)]$. The map at the near right of B3, chaotic orbits can exist on $J_2$ and $J_3$. The border is given by $B_3 = [a, d] F(D_4) = y_p$ where $y_p$ is an unstable fixed point as $y_p = \beta_s(y_p + 1) + 1 - d$ is satisfied. On this border, $d \approx 3.050$ for $a = 2.2$.

On border B4 in Fig. 5, we can see that chaotic orbits are changed into co-existence phenomena of chaos and complicated SSPOs. The return maps at the near left and right of B4 are shown in Fig. 6 (g) and (h), respectively. Fig. 6 (g) exhibits chaotic orbits on $J_2$ and $J_1$. As $d$ increases and passes through B4, $J_3$ is to include the flat segment as shown in Fig. 6 (h). In this case complicated SSPOs appear and $J_2$ remain an invariant interval. The border is given by $B_4 = [a, d] D_{\infty} = F(D_3))$. On this border, $d \approx 3.084$ for $a = 2.2$.

5. Conclusions

We have studied complicated superstable phenomena in a non-autonomous PWC circuit. PWC vector field defines the PWL trajectory and embedded return map is given explicitly. The circuit exhibits rich SSPOs and related bifurcation phenomena. Using the map, we have analyzed bifurcation phenomena for SSPOs precisely. Future problems include analysis for wider parameter space and consideration of engineering applications.

References

Codimension-2 bifurcation points organizing the bifurcation scenario of a hysteresis oscillator

Federico Bizzarri†, Oscar De Feo‡, Daniele Stellardo‡, Marco Storace‡

†Biophysical and Electronic Engineering Department, University of Genoa
Via Opera Pia 11a, I-16145 Genova, Italy
‡Microelectronic Engineering Dept., Faculty of Engineering, University College Cork
Enterprise Centre, North Mall Cork, Ireland

Email: Marco.Storace@unige.it

Abstract—In this paper we analyze some codimension-2 bifurcation points organizing the bifurcation scenario of a model of a hysteresis-based circuit oscillator. The proposed results have been obtained by resorting to numerical continuation methods. Such results largely justify the complexity of the scenario obtained by resorting to brute-force techniques.

1. Introduction

The hysteretic oscillator this paper deals with has been extensively studied in the last decade by means of several methods: brute force simulations [1], averaging techniques [2], and combinations of numerical continuation techniques and normal forms theory [3–6]. However, the variety of dynamical behaviors the oscillator exhibits has not yet been explained. In particular, it has not yet been found whether the richness of the observable dynamical behaviors can be justified by few characteristic nonlinear phenomena organizing the overall scenario in the parameter space. Here, we report the bifurcation analysis of some periodic solutions of an oscillator model. The results suggest a parameter space simply organized by characteristic codimension-2 bifurcation points.

2. The system

The normalized system of equations modeling the dynamics of the hysteresis oscillator can be written as follows:

\[
\begin{align*}
\dot{x}_1 &= -(x_1 + x_2) \\
\dot{x}_2 &= (2 + p_1)(x_1 + x_2) - x_2 - p_2 x_3 \\
\dot{x}_3 &= p_3 (\Psi - p_4 \sinh(x_3)) \\
x_1 &= x_3 = \sinh \left( \frac{x_3}{p_1} \right) + \Psi
\end{align*}
\]

(1)

where the bifurcation parameters are \( p_1 \) and \( p_2 \), whereas \( p_3 = 300, p_4 = 2.97 E - 24 \), and \( p_5 = 77.22 E - 12 \) are held fixed. The only nonlinear equation of the ODE system is Eq. (1c), where \( \Psi \) is implicitly defined by Eq. (1d).

From Eqs. (1), it is evident that (i) the equilibria of the system lie on the plane \( x_1 = -x_2 \), (ii) their positions depend on \( p_2 \) only, and (iii) the origin \( E_0 = (0, 0, 0) \) is a trivial equilibrium for any parameter values. Less evidently, the stability of each equilibrium point depends on both bifurcation parameters [3].

System (1) is \( \mathbb{Z}_2 \)-equivariant [5] under the transformation \( (x_1, x_2, x_3) \rightarrow (-x_1, -x_2, -x_3) \). In other words, there is a mirror symmetry with respect to \( E_0 \). As a consequence, pitchfork (or symmetry breaking) bifurcations break the \( \mathbb{Z}_2 \)-symmetry of fixed (i.e., invariant under the symmetry transformation) invariant sets and give rise to a symmetric pair of asymmetric invariant sets that are mirror image copies of each other. Henceforth, we denote as F-sets and S-sets the fixed and the symmetric invariant sets, respectively. Whilst \( E_0 \) is of F-type, in some regions of the parameter space there are other two S-type system equilibria, denoted as \( E_r \) and \( E_l \). In the following, \( E_r \) denotes both \( E_+ \) and \( E_0 \) (the same holds, mutatis mutandis, for other S-sets).

3. Observed families of periodic solutions

If we properly choose the values of the bifurcation parameters, we can observe different kind of periodic behaviors. Such behaviors can be naturally grouped into families indexed according to three integer numbers \( (n^+, n^-, n^0) \), representing the numbers of turns done by the limit cycle in the upper\(^1\), central, and lower region of the state-space, respectively. Examples, of both F- and S-type, of these periodic solutions are reported in Fig. 1.

Experimentally, the families of periodic solutions are organized in the parameter space as shown in Fig. 2, where the coloring is organized controlling the RGB components as functions of the \( n^+, n^0, \) and \( n^- \) indices, respectively. It is neither clear nor evident, how the families of solutions are linked to each other and how the chaos is organized by such periodic families.

\(^1\)Upper, lower, and central are meant with respect to the \( x_1 \) variable.
Figure 1: Families of periodic solutions in the state space \((x_1, x_2, x_3)\): (a) – upper S-type limit cycle \((4, 4, 0)\); (b) – F-type limit cycle \((3, 0, 3)\); (c) – large upper S-type limit cycle \((3, 3, 3)\); (d) – F-type limit cycle \((4, 2, 4)\).

Figure 2: Bifurcation scenario \(\ln(p_1)\) vs. \(\ln(p_2)\). Brute force scenario colored controlling the RGB components proportionally to the \(n^+, n^0\), and \(n^-\) indices; pale pink color corresponds to unbounded dynamics. Continuation curves colored as follows: cyan flips, red folds, green Hopfs, blue pitchforks, black heteroclinic and grey homoclinic.

4. Bifurcation analysis: Linking the families

The previously studied bifurcations [3] do not justify the presence of the families of limit cycles, how they are linked, and why they become chaotic. By using continuation techniques (AUTO2000 [4]), here we explain the organization of these families by means of few codimension-2 bifurcation points, linked by bifurcation curves. Since many of these points and curves (shown in Fig. 2) are very close to each other, they are sketched in Figs. 3 and 4.

We start from the limit cycle \((C_0)\) emerging from one of the previously studied bifurcations (the Hopf \(H_0\) of the origin \(E_0\)) and we see how few codimension-2 bifurcation points involving \(C_0\) justify all the families and the transitions between them. In the presented sketches, the dots denote unstable equilibria, with \(E_0\) blue and \(E_+\) green, the solid (dashed) closed lines denote stable (unstable) cycles, with \(C_0\) in blue.

4.1. Bogdanov-Takens: Existence of a global bifurcation

At very high values of \(p_1\), at the intersection \((\ln(p_1) \approx 17.58, \ln(p_2) \approx 0)\) between \(H_0\) and the pitchfork of the origin \(P_k\) [3] (see Fig. 3), \(E_0\) undergoes a Bogdanov-Takens (double zero) bifurcation (point \(BT\) in Fig. 3). Hence, because of the \(\mathbb{Z}_2\) symmetry of the system, from the intersection point \(BT\) a global bifurcation curve \(Ht\) roots, on which the cycle \(C_0\) degenerates to a double heteroclinic connection between \(E_+\) and \(E_-\).
However, due to the saddle (real) nature of $E_\pm$, $BT$ does not locally justify the existence of the periodic families.

### 4.2. Fold-Pitchfork: Organization of the symmetry breaking

The heteroclinic bifurcation, being global, is a good candidate for organizing more complex behaviors. Indeed, by continuing $Ht$ far away from $BT$, we discovered a degeneracy point $FP$, marking a codim-2 Fold-Pitchfork bifurcation for the limit cycle $C_0$. At this point, the already structurally unstable double heteroclinic connection undergoes a concurrent symmetry breaking. From this point, four further bifurcation curves root:

1. A pitchfork bifurcation ($PC$) of limit cycles, along which two S-type limit cycles ($C_\pm$, represented in green in the sketches) emerge from the F-type limit cycle $C_0$;

2. A fold bifurcation ($F_0$) of limit cycles, along which two F-type limit cycles ($C_0$ and, say, $C_h$, represented in magenta in the sketches) collide and disappear. Both cycles undergo to the heteroclinic bifurcation $Ht$, but in different ways.

3-4. A pair of coinciding homoclinic bifurcations $Hm$, where the two S-type limit cycles appearing at $P_C$ collide with the equilibria $E_\pm$, thus breaking and disappearing.

The aforesaid bifurcation lines are arranged around the point $FP$ as shown in Fig. 2 and sketched in Fig. 4, where ideas of the bifurcating invariant sets are given in the corresponding regions. Less evident is the existence of a period doubling bifurcation curve ($PD$) rooted in $FP$ and concerning the S-type limit cycles. This suggests the non-orientability of the flux around these cycles in the neighborhood of $FP$ and, consequently, a more complex scenario [8], which however we did not investigate further. Indeed, in this complex scenario, the still saddle (real) nature of the equilibria $E_\pm$ does not justify the existence of chaotic solutions.

### 4.3. Belyakov: Organization of chaotic solutions

Proceeding further, $Ht$ undergoes a Belyakov degeneracy, where the equilibria $E_\pm$ bearing the double heteroclinic connection change from saddle (real) to saddle-focus. The theory predicts the rooting from this point of several families, of infinite cardinality, of bifurcation curves [7; 8]. However, these curves accumulate exponentially on $Ht$ and the outer ones delimitate a region (light grey region in Fig. 4), where wild (chaotic) trajectories can be observed [9] according to the well-known Shil’nikov theorem [5]. Similarly, also the homoclinic curve $Hm$ rooted in $FP$ undergoes the same degeneracy and, consequently, a region where S-type wild trajectories can be observed takes place in the parameter space (dark grey region in Fig. 4). Concrete chaotic behaviors can be observed in the parameter space in correspondence of white and light blue regions in Fig. 2.

Whilst the Belyakov points justify the existence of chaotic solutions associated to both F- and S-type limit cycles, they do not explain the existence of different families of limit cycles characterized by different numbers $n^{+,-}$ of turns.

### 4.4. Shil’nikov-Hopf: Organization of the $n^{+,-}$–turns

Both $Ht$ and $Hm$ bifurcation lines end on the Hopf bifurcation curve ($H_\pm$) of the equilibria $E_\pm$ (Fig. 4). The ending points $SH_{1,2}$ are codimension-2 bifurcations known
as Shil’nikov-Hopf bifurcations. At this point, the homoclinic loop breaks into a double heteroclinic connection between a limit cycle and an equilibrium point [8] and, similarly, the double heteroclinic connection between equilibria breaks into a more complex multiple heteroclinic connection between equilibria and cycles. The theory predicts a very complex scenario nearby such a degeneracy [8], where infinite, non necessarily accumulating, bifurcations organize the existence of limit cycles characterized by different numbers of turns around the pair of limit cycles emerged on \( H_s \). Consequently, since the limit cycles appear in the upper (lower) region of the state-space, these degeneracies justify the existence of F-type (from \( SH_1 \) and \( Ht \)) as well as S-type (from \( SH_2 \) and \( Hm \)) limit cycles with different numbers \( n^{-+} \) of turns. However, no one of these bifurcation structures justifies the formation of the central \( (n^0) \) turns.

4.5. A blue-sky catastrophe: Formation of the \( n^0 \)-turns

The continuation of F-type limit cycles with respect to \( p_2 \), at values of \( p_1 \) below the level of the two \( SH \) points, usually propose a scenario as the one reported in Fig. 5 and corresponding to the yellow line in Fig. 2.

```
\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure5.pdf}
\caption{Typical blue-sky scenario for a F-type limit cycle \( (ln(p_1) = -1.2) \): (a) – Period vs. \( ln(p_2) \); (b) – Length vs. \( ln(p_2) \).}
\end{figure}
```

This characteristic scenario, with both period and length increasing towards infinity, suggests the existence of the so-called blue-sky catastrophe, where two limit cycles collide transversally, with the trajectory of one winding on the other. In our case, the trajectories tend to wind around the limit cycle \( C_0 \) (appeared from the Hopf of the origin \( E_0 \)), thus justifying the formation of the central \( (n^0) \) turns.

Unfortunately, blue-sky bifurcations cannot be numerically continued. Extensive simulations suggest the existence of infinite of them, one for each family of cycles with different \( n^{-+} \) turns. These bifurcation loci appear to be U-shaped (green points in Fig. 2, at the transitions from blue to grey regions in the “boomerangs” on the right side of the diagram), with the upper part converging towards the degeneracy \( FP \).

5. Conclusions

Simulations of the hysteresis-based oscillator model show the existence of several families of periodic and chaotic solutions characterized by different numbers of turns in the upper, central, and lower part of the state-space. The detailed bifurcation analysis reported here shows how their existence is justified, as well as organized in the parameter space, by few codimension-2 bifurcation points.

Acknowledgments

This work was supported by MIUR, within the FIRB framework (project RBNE012NSW), and by University of Genoa.

References


Study of hyperchaotic multi-scroll attractors via Adomian decomposition with application to synchronization and control

Donato Cafagna and Giuseppe Grassi

Dipartimento Ingegneria Innovazione, Università di Lecce
Via Monteroni, 73100 Lecce, Italy
Email: donato.cafagna@unile.it, giuseppe.grassi@unile.it

Abstract– In this paper the attention is focused on the numerical study of hyperchaotic 2D-scroll attractors via the Adomian decomposition method. The approach, which provides series solutions of the system equations, is first applied to weakly-coupled Chua’s circuits with Hermite interpolating polynomials. Then the method is utilized for achieving hyperchaos synchronization and control of the two coupled Chua’s circuits. The approach presents two main features, i.e., the system nonlinearity is preserved and the chaotic solution is provided in a closed form.

1. Introduction

In recent years the study of chaotic and hyperchaotic multi-scroll attractors has received increasing attention [1]-[2]. In this paper a further contribution to the numerical analysis of coupled Chua’s circuits generating 2D-scroll attractors is given. The starting point of the proposed approach is the Adomian decomposition method [3]. It represents a decomposition technique that doesn’t change the nonlinear physical problem into a convenient one for use with linear theory, but provides series solutions which generally converge very rapidly [3]. In particular, the Adomian decomposition provides immediate and visible symbolic terms of analytic solutions as well as numerical solutions to nonlinear differential equations. Moreover, differently from the Runge-Kutta method, the solution is found without linearization or discretization [3].

Note that, as far as the authors know, until now the Adomian decomposition has not been applied to synchronization and control of hyperchaotic circuits. Therefore, this paper exploits the Adomian method for analyzing the hyperchaotic dynamics of two weakly-coupled Chua’s circuits, with application to their synchronization and control. The paper is organized as follows. In Section 2 the basic notions of the Adomian decomposition are briefly illustrated. Moreover, the method is utilized for studying hyperchaotic coupled Chua’s circuits with a novel type of nonlinearity. The introduced nonlinearity, which belongs to the class of the Hermite polynomials, is well suited for calculating the Adomian polynomials required for the series solution of the decomposition approach. Based on these results, in Section 3 the Adomian approach is applied to hyperchaos synchronization and control of the two coupled Chua’s circuits. Numerical results demonstrate that the method is accurate in studying hyperchaotic dynamics and readily feasible in achieving both synchronization and control.

2. Study of chaotic dynamics via Adomian decomposition

2.1. Basic Notion

Given the functions \( \mathbf{x}(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T \) and \( \mathbf{g}(t) = (g_1(t), g_2(t), \ldots, g_n(t))^T \), consider the equation

\[
F(\mathbf{x}(t)) = \mathbf{g}(t)
\]

where \( F \) represents a nonlinear ordinary differential operator involving both linear and nonlinear terms. The Adomian decomposition requires that \( F \) be separated into three terms \( F = D + L + N \), where \( (D + L) \) form the linear vector and \( N \) is a nonlinear vector [3]. Here \( D \) is chosen to be easily invertible and \( L \) is the remainder of the linear term. Thus, system (1) can be written as

\[
D\mathbf{x} + L\mathbf{x} + N\mathbf{x} = \mathbf{g}(t)
\]

with \( \mathbf{x}(t_0) = \mathbf{x}_0 \). Let’s apply the inverse operator \( D^{-1} \) to the expression \( D\mathbf{x} = g(t) - L\mathbf{x} - N\mathbf{x} \), that is:

\[
\mathbf{x} = k + D^{-1}g(t) - D^{-1}L\mathbf{x} - D^{-1}N\mathbf{x}
\]

where \( k \) is the kernel of the operator \( D^{-1} \). Now, the vector solution \( \mathbf{x}(t) \) and the \( N\mathbf{x} \) term are written as:

\[
\mathbf{x} = \sum_{i=0}^{\infty} \mathbf{x}_i = \sum_{i=0}^{\infty} \mathbf{x}_1^i \ldots \mathbf{x}_n^i
\]

(4a)

\[
N\mathbf{x} = \sum_{i=0}^{\infty} \mathbf{A}_i = \sum_{i=0}^{\infty} \mathbf{A}_1^i \ldots \mathbf{A}_n^i
\]

(4b)

where the vector components of \( \mathbf{A}_i \) are the Adomian polynomials, whereas the superscript \( i \) indicates the element of the series. By using (4b), the elements \( x_i^k \) in (4a) can be recursively defined as follows:

\[
x_0^k = k + D^{-1}g(t)
\]

\[
x_1^k = -D^{-1}Lx_0^0 - D^{-1}A_1^0
\]

\[
x_2^k = -D^{-1}Lx_1^0 - D^{-1}A_1^1
\]

\[
\ldots
\]

\[
x_i^k = -D^{-1}Lx_{i-1}^0 - D^{-1}A_1^{i-1}
\]

(5)

where the \( j \)-th component of the polynomial \( A_1^i \) is [3]:

\[
A_1^i = \frac{1}{i!} \int \frac{d^i}{d\lambda^i} N(v_j'(\lambda)) d\lambda
\]

(6a)

\[
v_j'(\lambda) = \sum_{j=0}^i \frac{d}{d\lambda} x_j^i
\]

(6b)

with \( i = 0, 1, \ldots, \) and \( j = 1, \ldots, n \). Note that when a superscript is applied without round brackets it indicates the element of a series (for example, \( x_j^i \) in (6b)), otherwise it raises the corresponding power (for example, \( (\lambda)^j \) in (6b)). Moreover, note that (6) can be well adapted to calculate \( A_1^i \) by applying numerical packages like...
Matlab, Mathematica or Maple. Now, by substituting equations 4(a)-4(b) in equation (3), the infinite series solution of system (1) is:

\[ x = \sum_{i=0}^{\infty} x_i^I = k + D^{-1} g(t) - D^{-1} \left( \sum_{i=0}^{\infty} x_i^I \right) - D^{-1} \left( \sum_{i=0}^{\infty} A_i^I \right) \]  (7)

Finally, in order to obtain a practical solution of system (1) in the interval \([t_0, t]\), the infinite series (7) turns into the \((m+1)\)-terms series:

\[ \bar{x}(t) = \sum_{i=0}^{m} x_i^I + c_j^I (t-t_0)^2 + \frac{(t-t_0)^3}{3!} + \ldots + \frac{(t-t_0)^m}{m!}, \quad j = 1, \ldots, m \]  (8)

where the expressions of \(c_j^I\) depend on the parameters and the initial conditions of system (1). Differently from the Runge-Kutta method, the solution (8) has been obtained preserving the system nonlinearity, that is, without linear approximation. Note that, even though the solution (8) is a truncation of the infinite series solution (7) up to few terms, this is enough to give an accurate result. In fact, in [3] it has been shown that a practical value of \(m\) is four.

2.2. Analysis of Hyperchaotic 2D-scroll Attractors

Herein the attention is focused on the generation of hyperchaotic 3x4-scroll attractors in coupled Chua's circuits. The results in [2] demonstrate that such 3x4-scroll attractor can be generated via a sine-type nonlinearity, that is, a nonlinear function obtained by combining a sine function and a piece-wise linear function. Since this type of nonlinearity cannot be expanded by using the Adomian decomposition, herein it is shown that the 3x4-scroll attractor can be obtained by replacing the nonlinearities adopted in [2] with proper Hermite interpolating polynomials. To this purpose, the state equations of the coupled Chua's circuits are [4]:

\[
\begin{align*}
\dot{x}_1 &= \alpha \left[ x_2 - H_1(x_1) \right] \\
\dot{x}_2 &= x_1 - x_2 + x_3 + M(x_5 - x_3) \\
\dot{x}_3 &= -\beta x_2 \\
\dot{x}_4 &= \alpha \left[ x_1 - H_2(x_1) \right] \\
\dot{x}_5 &= x_4 - x_3 + x_6 + M(x_5 - x_3)
\end{align*}
\]  (9)

where \(\alpha = 10.814\), \(\beta = 14\), \(M = 0.25\), whereas \(H_1(\cdot)\) and \(H_2(\cdot)\) are the following Hermite polynomials:

\[
\begin{align*}
H_1(x_1) &= a_1(x_1)^3 + b_1(x_1)^{11} + c_1(x_1)^9 + d_1(x_1)^7 + e_1(x_1)^5 + f_1(x_1)^3 + g_1 x_1 \\
H_2(x_1) &= a_2(x_1)^3 + b_2(x_1)^{11} + c_2(x_1)^9 + d_2(x_1)^7 + e_2(x_1)^5 + f_2(x_1)^3 + g_2(x_1)^3 + h_2(x_1)^1 + i_2 x_1
\end{align*}
\]  (10a)

The parameters in (10) are reported in Table 1 and have been calculated with the aim to osculate the nonlinearities reported in [2] (see Fig.1 for the nonlinearity \(H_2(x_1)\)).

![Fig.1 The polynomial \(H_2(x_1)\) osculates the sine-type nonlinearity in [2] within the dynamic range of the 4-scroll in the second circuit.](image)

Note that the chosen Hermite polynomials have the same equilibrium points, with the same slopes, of the sine-type nonlinearities in [2]. Therefore, the equilibria of the proposed novel nonlinearities share the same dynamic properties of the equilibria of the sine-type nonlinearities. These considerations guarantee the existence of the hyperchaotic 3x4-scroll attractor [2]. By applying the method to the proposed coupled Chua’s circuits, the infinite series solution of (9) in the interval \([t_0, t]\) is:

\[
\begin{align*}
\begin{bmatrix}
\dot{x}_1(t) \\
\dot{x}_2(t) \\
\dot{x}_3(t) \\
\dot{x}_4(t) \\
\dot{x}_5(t)
\end{bmatrix} &= 
\begin{bmatrix}
\alpha x_2 \\
-\beta x_2 \\
\alpha x_1 \\
M x_4 + x_3 (1+M) + x_6 \\
-\beta x_1
\end{bmatrix} dt
\end{align*}
\]  (11)

By expanding the Hermite polynomials \(H_1(x_1)\) and \(H_2(x_1)\) via the formula (6), the expression of the five-term series solution of (11) in the generic subinterval \([t_k, t_{k+1}]\) is:

\[
\begin{align*}
H_1(x_1) &= a_1 x_1^3 + b_1 x_1^{11} + c_1 x_1^9 + d_1 x_1^7 + e_1 x_1^5 + f_1 x_1^3 + g_1 x_1 \\
H_2(x_1) &= a_2 x_1^3 + b_2 x_1^{11} + c_2 x_1^9 + d_2 x_1^7 + e_2 x_1^5 + f_2 x_1^3 + g_2 x_1^3 + h_2 x_1^1 + i_2 x_1
\end{align*}
\]  (10b)
The proposed theoretical approach is confirmed by the

\[ x_1 = c_1^0 + c_1^1 (t-t_0) + c_1^2 (t-t_0)^2 + c_1^3 (t-t_0)^3 + c_1^4 (t-t_0)^4 \]
\[ x_2 = c_2^0 + c_2^1 (t-t_0) + c_2^2 (t-t_0)^2 + c_2^3 (t-t_0)^3 + c_2^4 (t-t_0)^4 \]
\[ x_3 = c_3^0 + c_3^1 (t-t_0) + c_3^2 (t-t_0)^2 + c_3^3 (t-t_0)^3 + c_3^4 (t-t_0)^4 \]
\[ x_4 = c_4^0 + c_4^1 (t-t_0) + c_4^2 (t-t_0)^2 + c_4^3 (t-t_0)^3 + c_4^4 (t-t_0)^4 \]

\[ (12) \]

where the coefficients \( c_i \) in the subintervals \([t_i, t_{i+1}]\) have been obtained by exploiting the computer-based algorithm in [5]. Note that the solution of (9) is a composite of Adomian solutions (12) over the subintervals \([t_i, t_{i+1}]\), which have been spliced (concatenated) together. Such solution generates the attractors reported in Fig.2. For example, by considering the initial conditions \((x_i(0), x_{i+1}(0), x_{i+2}(0), x_{i+3}(0), x_{i+4}(0)) = (0.1, 0.2, 0.1, 0.1, 0)\)

\[ (13) \]

the numerical expression of (12) in the interval \([0, 0.1]\) is:
\[ \tilde{x}_1(t) = 0.1 + 2.0194t - 2.7925t^2 + 0.6488t^3 - 0.5263t^4 \]
\[ \tilde{x}_2(t) = 0.2 - 0.2500t - 0.2278t^2 - 0.2638t^3 + 0.5089t^4 \]
\[ \tilde{x}_3(t) = -0.1 - 2.8000t + 1.7500t^2 + 1.0630t^3 + 0.9232t^4 \]
\[ \tilde{x}_4(t) = -0.1 - 0.1434t + 0.1680t^2 - 0.4031t^3 - 0.2089t^4 \]
\[ (14) \]

\[ \tilde{x}_5(t) = 0.0500t + 0.1342t^2 - 0.0237t^3 + 0.0467t^4 \]
\[ \tilde{x}_6(t) = 0.1 - 0.3500t^2 + 0.6262t^3 + 0.0830t^4 \]

Figure 2 and equation (14) highlight that the method is an effective tool for studying complex behaviors, providing a different insight into the system dynamics.

**3. Application to synchronization and control**

This Section aims to show that the Adomian decomposition provides a different point of view to both chaos synchronization and control.

**3.1 Synchronization**

The attention is now focused on the synchronization of the hyperchaotic 2D-scroll attractors generated by system (9). By considering the drive system in the form (12), the equations of the uncontrolled response system are:
\[ \dot{y}_1 = d_1^0 + d_1^1 (t-t_0) + d_1^2 (t-t_0)^2 + d_1^3 (t-t_0)^3 + d_1^4 (t-t_0)^4 \]
\[ \dot{y}_2 = d_2^0 + d_2^1 (t-t_0) + d_2^2 (t-t_0)^2 + d_2^3 (t-t_0)^3 + d_2^4 (t-t_0)^4 \]
\[ \dot{y}_3 = d_3^0 + d_3^1 (t-t_0) + d_3^2 (t-t_0)^2 + d_3^3 (t-t_0)^3 + d_3^4 (t-t_0)^4 \]
\[ \dot{y}_4 = d_4^0 + d_4^1 (t-t_0) + d_4^2 (t-t_0)^2 + d_4^3 (t-t_0)^3 + d_4^4 (t-t_0)^4 \]
\[ (15) \]

where the expressions of \( d_i \) depend on the initial conditions of the response system. It is well known that the drive system (12) and the response system (15) are synchronized when \( e_i(t) = (\tilde{x}_i(t) - \tilde{y}_i(t)) \rightarrow 0 \) for \( t \rightarrow \infty \) \((i = 1...6)\). To this purpose, a proper coupling between the drive and response systems has to be designed. In the proposed case, the synchronization signals \( w(t) \) are functions of the time as reported in the following equation
\[ w(t) = \left\{ \begin{array}{ll}
0.0500t & \text{if } t \leq 0.1 \\
0.1 & \text{if } 0.1 < t \leq 0.2 \\
0.2 & \text{if } 0.2 < t \leq 0.3 \\
0.3 & \text{if } 0.3 < t \leq 0.4 \\
0.4 & \text{if } 0.4 < t \leq 0.5 \\
0.5 & \text{if } 0.5 < t \leq 0.6 \\
0.6 & \text{if } 0.6 < t \\
\end{array} \right. \]

\[ (16) \]

By focusing the attention on (12) and (16), it can be argued that the method provides a straightforward way for obtaining synchronization. Namely, it can be readily demonstrated that the error between (12) and (16) asymptotically approaches zero if the signals \( w(t) \) are:
\[ \left\{ \begin{array}{l}
\dot{w}_0(t) = c_0^1 - d_0^1 (t-t_0) + 2c_0^2 (t-t_0)^2 - 6c_0^3 (t-t_0)^3 + 24c_0^4 (t-t_0)^4 \\
\dot{w}_1(t) = c_1^1 - d_1^1 (t-t_0) + 2c_1^2 (t-t_0)^2 - 6c_1^3 (t-t_0)^3 + 24c_1^4 (t-t_0)^4 \\
\dot{w}_2(t) = c_2^1 - d_2^1 (t-t_0) + 2c_2^2 (t-t_0)^2 - 6c_2^3 (t-t_0)^3 + 24c_2^4 (t-t_0)^4 \\
\dot{w}_3(t) = c_3^1 - d_3^1 (t-t_0) + 2c_3^2 (t-t_0)^2 - 6c_3^3 (t-t_0)^3 + 24c_3^4 (t-t_0)^4 \\
\dot{w}_4(t) = c_4^1 - d_4^1 (t-t_0) + 2c_4^2 (t-t_0)^2 - 6c_4^3 (t-t_0)^3 + 24c_4^4 (t-t_0)^4 \\
\dot{w}_5(t) = c_5^1 - d_5^1 (t-t_0) + 2c_5^2 (t-t_0)^2 - 6c_5^3 (t-t_0)^3 + 24c_5^4 (t-t_0)^4 \\
\dot{w}_6(t) = c_6^1 - d_6^1 (t-t_0) + 2c_6^2 (t-t_0)^2 - 6c_6^3 (t-t_0)^3 + 24c_6^4 (t-t_0)^4 \\
\end{array} \right. \]

\[ (17) \]
simulation results reported in Fig. 3, which shows the time behavior of the state variables $\dot{x}_1$ and $\dot{y}_1$. 

![Fig.3 Hyperchaotic synchronization: time behavior of $\dot{x}_1$ (dot line) and $\dot{y}_1$. The synchronization signal is turned on at $t = 100$.](image)

3.2 Control

Herein a controller obtained via the Adomian method is used for stabilizing stable and unstable equilibrium points of (9). Let $x_{eq} = (x_{1eq}, x_{2eq}, x_{3eq}, x_{4eq})^T$ be an equilibrium point of the uncontrolled system (9). The objective is to find a controller $u(t)$ such that the state of the controlled system in the form

$$
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
\dot{x}_3 \\
\dot{x}_4
\end{bmatrix} = \begin{bmatrix}
c_1' + c_1'(t-t_e) + c_1'(t-t_e)^2 + c_1'(t-t_e)^3 \\
c_2' + c_2'(t-t_e) + c_2'(t-t_e)^2 + c_2'(t-t_e)^3 \\
c_3' + c_3'(t-t_e) + c_3'(t-t_e)^2 + c_3'(t-t_e)^3 \\
c_4' + c_4'(t-t_e) + c_4'(t-t_e)^2 + c_4'(t-t_e)^3
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{bmatrix}
$$

(18)

is stabilized to $x_{eq}$. By looking at equations (18), it can be argued that the Adomian decomposition provides a straightforward way for obtaining the controller $u(t)$. Namely, it can be readily shown that if the controller is:

$$
\begin{align*}
&u_1(t) = c_1' - \frac{c_1'(t-t_e)}{2} + \frac{c_1'(t-t_e)^2}{6} + \frac{c_1'(t-t_e)^3}{24} \\
&u_2(t) = c_2' - \frac{c_2'(t-t_e)}{2} + \frac{c_2'(t-t_e)^2}{6} + \frac{c_2'(t-t_e)^3}{24} \\
&u_3(t) = c_3' - \frac{c_3'(t-t_e)}{2} + \frac{c_3'(t-t_e)^2}{6} + \frac{c_3'(t-t_e)^3}{24} \\
&u_4(t) = c_4' - \frac{c_4'(t-t_e)}{2} + \frac{c_4'(t-t_e)^2}{6} + \frac{c_4'(t-t_e)^3}{24}
\end{align*}
$$

(19)

the controlled system is stabilized to any equilibrium point $x_{eq}$. In order to show how the method works, the attention is focused on the stabilization problem of the unstable saddle type-II equilibria [6]. Figure 4 shows that the considered system is stabilized to the unstable equilibrium point $x_{eq} = (2.6, 0, -2.6, 0, 0)^T$, confirming the effectiveness of the Adomian-based control method.

![Fig.4 Hyperchaotic control: stabilization of $\dot{x}_1$ to $x_{eq} = 2.6$. The control signal is turned on at $t = 100$. Similar behaviors have been obtained for the remaining state variables.](image)

4. Conclusions

By applying the Adomian method, this work has focused on the study of two coupled Chua’s circuits. The method, which provides series solutions of the system equations, has been applied by taking Hermite polynomials in the Chua’s circuits. The application of the method to the numerical analysis of hyperchaotic circuits presents two advantages: i) the nonlinearity is preserved, which can be of great importance in the analysis of complex dynamics; ii) the solution is provided in a closed form, which can give an additional insight into the system dynamics. Based on the latter feature, the method is applied to synchronization and control of the two coupled Chua’s circuits. Finally, we would remark that the capability of the method in providing the solution in a closed form leads to a different point of view to the analysis, synchronization and control of chaotic dynamics.

References


Generating Chaos with a Linear Filter

Ned J. Corron, Scott T. Hayes, Shawn D. Pethel, and Jonathan N. Blakely

U. S. Army RDECOM
AMSRD-AMR-WS-ST, Redstone Arsenal, Alabama 35898, USA
Email: ned.corron@us.army.mil

Abstract– Recently, it has been shown that chaos can be synthesized by the linear superposition of certain pulse basis functions. Here, we extend this result and show that a linear, second-order filter driven by a random signal can generate a waveform that is chaotic under time reversal. That is, the waveform exhibits determinism and a positive Lyapunov exponent when viewed backward in time. We demonstrate the filter using a passive electronic circuit, and the resulting waveform exhibits a Lorenz-like butterfly structure. This method for generating chaotic waveforms may be useful for a number of potential applications. The filter also demonstrates that chaos may be connected to physical theories beyond those described by a deterministic nonlinear dynamical system.

1. Introduction

Chaotic waveforms have been suggested for a number of applications, including communication and remote sensing. For these applications, one usually assumes a nonlinear dynamical system is required to generate the chaotic waveform. However, it has recently been shown that chaos can also be constructed by linear superposition of special basis pulses [1-4]. This surprising result implies that chaos can also be formally generated using a linear filter. However, such linear synthesis of chaos does not appear to be practical for a physical system. The pulse basis functions contain an infinitely long, exponentially increasing oscillation that culminates in a large central pulse and monotonic exponential decay. Thus, the exact filter is necessarily acausal and impractical to realize.

In this paper, we exploit the fact that the basis pulses do not need to be acausal when viewed in reverse time and develop a very simple system for generating chaotic waveforms. In particular, we find that a linear, second-order filter driven by a random bipolar signal can generate a waveform that is chaotic in reverse time. We call such dynamics reverse-time chaos. That is, when viewed backward in time, the waveform exhibits the essential qualities of a chaotic waveform, including determinism and a positive Lyapunov exponent. We implement the filter in an electronic circuit using only a few passive linear components, and we obtain a waveform that exhibits a Lorenz-like butterfly structure.

From previous work, it is known that a linear filter can increase the apparent dimensionality of a chaotic signal [5]. But here we show that a linear filter driven by a random process produces a waveform that appears to have been produced by a low-dimensional chaotic system. This phenomenon suggests that chaos may be connected to physical theories whose underlying framework is not that of a traditional deterministic nonlinear dynamical system. Furthermore, the simplicity of the filter mechanism implies one must allow for the possibility of reverse-time chaos occurring naturally. For example, efforts to detect determinism from randomness in a physical system may require the arrow of time to differentiate chaos from linear filtering of noise. Powerful algorithms that rely only on geometric analysis of state space structures, such as fractal dimension, template analysis [6], or false nearest neighbors [7], may not detect a difference between forward and reverse-time chaos.

Several potential technology applications exist for reverse-time chaos. For communications, symbolic dynamics can be encoded directly by modulating the polarity of the basis pulses [8]. At the receiver, the determinism of reverse-time chaos provides a form of intentional inter-symbol interference that can be processed using simple predictive filters [2, 9]. For correlation-based ranging using chaotic waveforms, we note that reverse-time waveforms will work equally well as forward-time chaos [10-13]. Reverse-time chaos may also be relevant to the prediction and control of electromagnetic interference in circuits driven by radio frequency signals [14, 15].

2. Reverse-Time Chaos

To demonstrate reverse-time chaos in a linear filter, we consider the driven second-order linear system

\[ \ddot{x} + 2\beta \dot{x} + \omega^2 x = s(t) \]  

where \( x(t) \) is the scalar state, \( \beta = \ln 2 \) is the decay rate, and \( \omega = 2\pi \) is the frequency of the damped oscillations. The input signal \( s(t) \) is

\[ s(t) = A \cdot s_n, \quad n \leq t < n + 1 \]  

where each \( s_n \in \{-1, 1\} \) is random and \( A \) is a fixed amplitude. Since equation (1) is linear, we set \( A = 1 \) without loss of generality. The homogeneous solution of equation (1) is

\[ x_h(t) = C \cdot 2^{-\frac{t}{\beta}} \cos(2\pi + \phi) \]
where \( C \) and \( \phi \) are integration constants. It is easy to show that the homogenous solution satisfies

\[
x_h(t+1) = \frac{1}{2} x_h(t)
\]

for all \( t \geq 0 \); in fact, requiring a solution that satisfies (4) defines the particular values we use here for the system parameters \( \beta \) and \( \omega \). [4]

To solve for a particular solution, we consider the response of the linear system to excitation by a unit pulse. That is, we solve the initial value problem

\[
\begin{cases}
x(0) = 0, & t = 0, 1 \leq t < 1 \\
x(0) = 0, & t \geq 1
\end{cases}
\]

subject to the homogenous initial conditions

\[
\zeta(0) = \tilde{\zeta}(0) = 0.
\]

We extend this solution to negative time by defining \( \zeta(t) = 0 \) for \( t < 0 \), and the complete pulse response is plotted in Fig. 1. A general solution to equation (1) is then found by superposition

\[
x(t) = \sum_{n=-\infty}^{\infty} s_n \zeta(t - n).
\]

Using the unit pulse response (6), we find

\[
x(t) = \frac{1}{4\pi^2 + (\ln2)^2} \left[ \begin{array}{cc}
1 - \exp\left\{ -2^{-t} \cos(2\pi) + \frac{\ln2}{2\pi} \sin(2\pi) \right\} & 0 \leq t < 1 \\
2^{-t} \cos(2\pi) + \frac{\ln2}{2\pi} \sin(2\pi) & t \geq 1
\end{array} \right]
\]

(8)

where the notation \([t]\) indicates the largest integer less than or equal to \( t \).

We claim the waveform (8) is chaotic when viewed in reverse time. This claim can be justified using the results of Hirata and Judd [3], who derived necessary conditions for the basis pulse function to assure the superposed dynamics are conjugate to a chaotic shift map. Instead, here we directly show that a shift dynamics exists in the general solution (8). To this end, we define a Poincaré return at integer times \( t = [t] \). The \( n \)th return is then

\[
x(n) = \frac{1}{4\pi^2 + (\ln2)^2} \sum_{i=1}^{\infty} 2^{-i} s_{n-i}.
\]

Defining the scaled return

\[
y_n = \frac{4\pi^2 + (\ln2)^2}{2} x(n) + \frac{1}{2}
\]

yields

\[
y_n = \sum_{i=1}^{\infty} 2^{-i} \sigma_{n-i}
\]

(11)

where \( \sigma_n = \frac{s_n + 1}{2} \) maps the bipolar symbols \( s_n \in \{-1, +1\} \) to the bits \( \sigma_n \in \{0, 1\} \). It is easy to verify that successive scaled returns (11) satisfy

\[
y_n = 2y_{n+1} \mod 1
\]

(12)

which is a chaotic shift map backward in time. Consequently, the continuous-time waveform (8) exhibits chaos in reverse time. We note this does not imply that integrating equation (1) backward in time will generate a chaotic waveform. Instead, this result implies that the general solution will, when viewed in reverse time, exhibit the properties of a “chaotic” waveform, namely...
FIG. 4. Driven RLC filter circuit for demonstrating reverse-time chaos.

determinism (the shift map) and a positive Lyapunov exponent ($\beta = \ln 2$).

In Fig. 2, we show a solution obtained by integrating equation (1) for a random sequence $s(t)$. The top plot shows the driving sequence and waveform in forward time. The bottom plot shows the same waveform in reverse time. The reverse-time waveform is similar to chaotic waveforms produced by the Lorenz system [16]. In Fig. 3 we show an “attractor” constructed using delay embedding of the waveform with $\Delta t = 1/3$. We note a similarity of the “attractor” in Fig. 3 with certain projections of the butterfly attractor of the Lorenz system.

3. Electronic Filter Circuit

Significantly, equation (1) models a number of physical systems including, for example, a damped linear pendulum. Here we demonstrate reverse-time chaos using the electronic filter shown in Fig. 4. Among the first circuits known to students, it is modeled as

$$LC \frac{d^2v}{dt^2} + RC \frac{dv}{dt} + v = v_{in}(t)$$

(13)

where $v(t)$ is the voltage across the capacitor. The applied drive voltage $v_{in}(t)$ is

$$v_{in}(t) = V_{in} \cdot s_n, \quad n \leq \frac{t}{T} < n + 1$$

(14)

where $V_{in}$ is a fixed amplitude, $s_n \in [-1,1]$ is a random sequence, and $T$ is the fundamental period. Introducing $\tau = \frac{t}{T}$ yields

$$\frac{d^2v}{d\tau^2} + \frac{TR}{L} \frac{dv}{d\tau} + \frac{T^2}{LC} v = f(\tau)$$

(15)

where

$$f(\tau) = \frac{T^2}{LC} V_{in} \cdot s_n, \quad n \leq \tau < n + 1.$$ 

(16)

We note that the system (15)-(16) is in the same form as system (1)-(2); thus, the circuit can exhibit reverse-time chaos for parameters satisfying

$$\frac{TR}{L} = 2 \ln 2$$

(17)

and

$$\frac{T^2}{LC} = 4\pi^2 + (\ln 2)^2.$$ 

(18)

We implement the circuit shown in Fig. 4 using discrete electronic components. We use an inductor with $L = 7.5$ mH and intrinsic series resistance $R = 57$ $\Omega$. The requirements (17) and (18) then give $T = 180$ $\mu s$ and $C = 0.11$ $\mu F$, respectively. To drive the circuit, we use a DSP card (Innovative Integration ADC64) hosted in a PC to generate random $\pm 1$-V bipolar bits at 5.6 kHz. The output waveform $v(t)$ is sampled at 100 kHz using a 12-bit data acquisition card (Keithley DAS-1802).

A typical output waveform captured from the filter is shown in Fig. 5. The top plot shows the input drive signal and output voltage in forward time. The bottom plot shows the same output waveform in reverse time. In Fig. 6 we show the measured “attractor” constructed by delay embedding with $\Delta t = 60$ $\mu s$. Again, we note the similarity of the reverse-time waveform and “attractor” to those of the chaotic Lorenz system.
4. Determinism and Randomness

The construction of such a simple passive electronic filter that generates reverse-time chaos when driven by a random signal indicates that such phenomena can easily occur in physical systems. Besides allowing possible technological applications, this capability impacts our view of determinism and randomness in nature. From an information point of view, deterministic chaos produces information by amplifying microscopic details of the initial state that were beyond the observer’s ability to measure [17]. Even though the unfolding information is new to the observer, the future dynamics are completely determined by the present state. In this way, chaos beautifully reconciles chance and determinism.

Reverse-time chaos turns determinism around: the present state stores the system’s entire history. The infinite sequence of prior random pulses can be reconstituted from a perfect knowledge of the present state of the filter. This property—that the present state determines the past—implies determinism when viewed in reverse time. Even though the information used to generate reverse-time chaos is truly random, this information is deterministically mapped to microscopic levels in the system state, eventually beyond the reach of any observer’s precision. Seeing further into the past requires ever greater measurement precision, implying a sensitive dependence on initial conditions and a positive Lyapunov exponent in reverse time. In this way, the random reverse-time waveform is entirely consistent with a chaotic, deterministic forward-time waveform.

5. Conclusion

In this paper, we have shown that a reverse-time chaotic waveform can be generated by a linear, second-order filter driven by a random source. We believe it is surprising that chaos in any form can be generated by such a remarkably simple system. This alternative method for generating broadband chaotic waveforms may be useful in a number of potential technological applications currently being considered. That the language of deterministic chaos provides a meaningful description for signals not generated by a nonlinear dynamical system suggests chaos may be more fundamental than previously supposed. At the very least, the possibility of chaos by such a simple mechanism provides a new perspective on the interplay and reconciliation of deterministic and random processes in nature.

References

Model-Following Control for Nonlinear Plants with Time Delay

Yoichi Ishikawa† and Yoshihisa Ishida †
†Department of Electronics and Communications, Meiji University
1-1-1 Higashi-mita, Tama-ku, Kawasaki 214-8571, Japan
Email: youichi76best@msn.com, ishida@isc.meiji.ac.jp

Abstract—In this paper, a design of control systems for nonlinear plants with time delay is proposed. It is derived that the control systems of any nonlinear plants with time delay can be reduced to a simple second-order model. For the identification of nonlinear plants, radial basis function neural networks, which are known for their stable learning capability and fast training, are used. In the simulation study of nonlinear plants with time delay, it is observed that the error between the plant output and the reference model output is negligibly small and shown that the proposed method yields a stable tracking performance.

1. Introduction

For nonlinear plants, fuzzy controllers [1] or the controllers using multi-layer perceptron neural networks (MLP-NNs) with back propagation algorithm [2] - [3] have been proposed. In fuzzy control, the advantages are that it is not necessary to identify any transfer function of the system (in advance) and that the input-output mapping can be easily represented. However, this approach inherently has the disadvantage in that the smart tuning of the membership function is still required. Although, the controllers using MLP-NNs, in contrast, can be adaptively tune the nonlinear input-output relations to a certain extent, the controllers so designed generally require long and/or iterative training of the MLP-NN parameters. Moreover, it is well known that they quite often suffer from numerical instability, due to the network parameters remaining at local minima during the training.

On the other hand, a family of radial basis function neural networks (RBF-NNs) have been applied to adaptive control [4] - [5]. It is well known that the RBF-NNs are stable as compared with the conventional MLP-NNs and a number of non-iterative network parameter tuning paradigms have also been proposed [6].

In this paper, we propose a design method for the model-following controllers using RBF-NNs. It has been shown that the feedback part of the difference equation can be estimated by the RBF-NNs. In the next section, the principle for designing such controllers is given. In Section 3, the design method for nonlinear plants is described. Section 4 is devoted to the simulation study. Finally, conclusions are remarked in Section 5.

2. Principle of the Design

2.1. Compensation of time delay

Let the difference equation of a n-order plant with time delay be given as

\[ y_p(k+1) = f[y_p(k), y_p(k-1), \ldots, y_p(k-n+1)] + \sum_{j=0}^{n-1} b_j u(k-j-L) \]  

(1)

where \( y_p(k) \) is the plant output, \( u(k) \) is the plant input, \( f[\cdot] \) is a nonlinear function, \( L \) is time delay, and \( b_0 \neq 0 \). By replacing \( k \) with \( k+1 \) in (1), we have

\[ y_p(k+2) = f[y_p(k+1), y_p(k), \ldots, y_p(k-n+2)] + \sum_{j=0}^{n-1} b_j u(k-j-L+1). \]  

(2)

In the above, when \( y_p(k+1) \) is set to \( y_p^{(1)}(k) \), we obtain

\[ y_p^{(1)}(k+1) = f[y_p^{(1)}(k), y_p^{(1)}(k-1), \ldots, y_p^{(1)}(k-n+1)] + \sum_{j=0}^{n-1} b_j u(k-j-L+1) \]  

(3)

where \( y_p^{(1)}(k) \) on the right hand side in (3) represents the prediction value \( y_p(k+1) \) at time \( k+1 \). Moreover, by replacing \( k \) with \( k+1 \) in (3), we obtain

\[ y_p^{(2)}(k+1) = f[y_p^{(1)}(k+1), y_p^{(1)}(k), \ldots, y_p^{(1)}(k-n+2)] + \sum_{j=0}^{n-1} b_j u(k-j-L+2) \]  

\[ = f[y_p^{(2)}(k), y_p^{(1)}(k-1), \ldots, y_p^{(1)}(k-n+1)] + \sum_{j=0}^{n-1} b_j u(k-j-L+2). \]  

(4)

By repeating the same procedure, the difference equation of a n-order plant without time delay can be obtained as

\[ y_p^{(n)}(k+1) = f[y_p^{(n)}(k), y_p^{(n)}(k-1), \ldots, y_p^{(n)}(k-n+1)] + \sum_{j=0}^{n-1} b_j u(k-j). \]  

(5)
2.2. The Design of Model-Following Controller

Next, let us assume that the difference equation for the reference model is given by a second-order system

\[ y_m(k + 1) = -a_{m1}y_m(k) - a_{m2}y_m(k - 1) + b_{m0}r(k) + b_{m1}r(k - 1) \]  

(6)

where \( y_m(k) \) is the model output and \( r(k) \) is the reference input. Then, the error \( e(k) \) between the model output and the plant output is defined as

\[ e(k) = y_m(k) - y_p^{(L)}(k). \]  

(7)

Now, let us assume that at the steady the following condition is satisfied:

\[ e(k + 1) = e(k). \]  

(8)

In this case, if \( e(0) \neq 0 \), \( \lim_{k \to \infty} e(k) \neq 0 \). Moreover, the stability of the controlled system may not be guaranteed. It will be justified later in this paper that these can be avoid. From (8)

\[ y_m(k + 1) - y_p^{(L)}(k + 1) = y_m(k) - y_p^{(L)}(k). \]  

(9)

By substituting (5) and (6) into (9) and replacing \( y_m(k) \) with \( y_p^{(L)}(k) \), we have

\[-a_{m1}y_p^{(L)}(k) - a_{m2}y_p^{(L)}(k - 1) + b_{m0}r(k) + b_{m1}r(k - 1) - f[k] - b_{au}u(k) - \sum_{j=1}^{n-1} b_j u(k - j) = y_p^{(L)}(k) \]

\[-f[k - 1] - b_{au}u(k - 1) - \sum_{j=1}^{n-1} b_j u(k - j - 1) \]  

(10)

where \( f[k] = \{ f[y_p^{(L)}(k), y_p^{(L)}(k - 1), \ldots, y_p^{(L)}(k - n + 1)] \}. \)

Therefore, a new relation for the plant input \( u(k) \) can be derived:

\[ u(k) = u(k - 1) + \frac{1}{b_0} \left\{ -(a_{m1} + 1)y_p^{(L)}(k) - a_{m2}y_p^{(L)}(k - 1) + b_{m0}r(k) + b_{m1}r(k - 1) - f[k] + f[k - 1] - \sum_{j=1}^{n-1} b_j [u(k - j) - u(k - j - 1)] \right\}. \]  

(11)

Although the proposed design method is extremely simple as compared with Narendra’s method [7], the same result is obtained. On the other hand, substituting (11) into (5), we obtain

\[ y_p^{(L)}(k + 1) = -a_{m1}y_p^{(L)}(k) - a_{m2}y_p^{(L)}(k - 1) + b_{m0}r(k) + b_{m1}r(k - 1). \]  

(12)

Then, the error \( e(k + 1) \) is given by

\[ e(k + 1) = -a_{m1}e(k) - a_{m2}e(k - 1). \]  

(13)

Therefore, the following condition is satisfied for arbitrary initial value \( e(0) \):

\[ \lim_{k \to \infty} e(k + 1) = \lim_{k \to \infty} e(k) = 0. \]  

(14)

Finally, we discuss the stability of the proposed method. Generally, since it is hard to analyze the stability of nonlinear control systems rigorously, we consider the following relation: In (11), let

\[ v(k + 1) \overset{\text{def}}{=} -(a_{m1} + 1)y_p^{(L)}(k) - a_{m2}y_p^{(L)}(k - 1) + b_{m0}r(k) + b_{m1}r(k - 1) - f[k] + f[k - 1]. \]  

(15)

Taking the z-transform for (11) gives

\[ \frac{U(z)}{V(z)} = \frac{1}{1 - z^{-1}} \sum_{j=1}^{n-1} b_j z^{-j}. \]  

(16)

In order to be stable for (16), all the zeros of the plants to be controlled must be located within a unit circle. Therefore, the proposed method is stable if the plants do not have any unstable zeros. Although this is also clear from some simulation experiments, we would like to investigate the stability of the proposed method in detail at another opportunity.

By the way, the proposed design method suffers from any load disturbance since the predictor can not estimate it. To solve this, the error between the plant output and the predictor output is fed back through the PID controller.

3. The Design for Nonlinear Plants

In system identification of nonlinear plants, the multi-layer perceptron neural networks (MLP-NNs) with back-propagation algorithm have been widely used. However, as discussed early, it is well known that employing the MLP-NNs involves rather heavy computation due to the iteratively updating of the weight vectors and the fact that there may always be a danger of the network parameters remaining at local minima. This limits their practical use, since an on-line processing is inevitable for controlling the dynamical systems. In this paper, we thus propose to use radial basis function neural networks (RBF-NNs) for representing the nonlinearity within the system identification.

3.1. Radial basis Function Neural Networks

Figure 1 illustrates an RBF-NN with \( N_i \) inputs, \( N_o \) radial basis functions (RBFs), and a single output. In Figure 1, each RBF is given by so-called a Gaussian response function:

\[ h_i = \exp\left( -\frac{\| x - c_i \|^2}{2\sigma^2} \right) \]  

(17)
where \( x = [x_1, \ldots, x_N]^T \) is the input vector, \( c_i \) is the centroid vector of the \( i \)-th RBF, \( \sigma \) is the radius, and \( || \cdot ||^2 \) denotes the \( L_2 \) norm. The output \( y \) of the RBF-NN is then given as the weighted linear sum of \( h_i \), i.e.,
\[
y = \frac{1}{\delta} \sum_{i=1}^{N_h} \theta_i w_i,
\]
(18)
where \( \delta \) is a constant and \( w = [w_1, w_2, \ldots, w_N]^T \) is called as the weight vector.

**Inputs**  
\( x_1 \)  
\( x_2 \)  
\( \vdots \)  
\( x_N \)

**Hidden Layer**  
\( h_1 \)  
\( h_2 \)  
\( \vdots \)  
\( h_N \)

**Output**  
\( y \)

1/\( \delta \)

**Figure 1: An RBF-NN.**

### 3.2. Nonlinear Plants

In the design, the function \( f[k] \) in (11) is replaced with the output of the RBF-NN and the input vector \( x(k) = [y_p(k), y_p(k-1), \ldots, y_p(k-n+1)]^T \) (i.e., \( N_l = n \)), that is
\[
f[k] = f[x^T(k)] = \frac{1}{\delta} \sum_{i=1}^{N_h} h_i(k) w_i
\]
with the response function
\[
h_i(k) = \exp \left( -\frac{||x(k) - c_i||^2}{2\sigma^2} \right).
\]
(20)

Then, the difference equation for nonlinear plants with time delay yields
\[
y_p(k + 1) = \frac{1}{\delta} \sum_{i=1}^{N_h} h_i(k) w_i + \sum_{j=0}^{n-1} b_j u(k - j - L)
\]
(21)
where \( L \) is time delay. Finally, the plant input \( u(k) \) is given by (11) with a nonlinear function (19).

### 3.3. Estimating Feedforward and Weight Parameters

To estimate both the feedforward parameters \( b_j \) and the weight parameters \( w_i \) of the RBF-NN in (21), we apply the singular value decomposition (SVD) to tuples of the set instead of applying the conventional least squares. Namely, the following relation is considered:
\[
AW = Y
\]
(22)
with
\[
A = \begin{bmatrix}
  u(k - N + 1) & u(k - N) & \cdots & u(k - N - n + 1) \\
  u(k - N + 2) & u(k - N + 1) & \cdots & u(k - N - n + 2) \\
  \vdots & \vdots & \ddots & \vdots \\
  u(k) & u(k - 1) & \cdots & u(k - n)
\end{bmatrix}
\]
\[
W = [b_0, b_1, \ldots, b_{n-1}, w_1, w_2, \ldots, w_N]^T.
\]
\[
Y = [y_p(k + 1), y_p(k - N + 2), \ldots, y_p(k + 1)]^T.
\]
where \( u(k) = 0 \) for \( k < 0 \). Applying the SVD, \( A \) can be rewritten as
\[
A = U S V^T
\]
(23)
where \( U = [u_1, u_2, \ldots, u_N] \) and \( V = [v_1, v_2, \ldots, v_{n+N_h}] \) are orthogonal to each other such that \( U^T V = I_{n+N_h} \) and \( V^T V = I_{n+N_h} \), respectively, and
\[
S = \begin{bmatrix}
  \Sigma & 0 \\
  0 & 0
\end{bmatrix}
\]
(24)
where \( \Sigma = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \) and \( \gamma < n + N_h \). Finally, \( W \) can be estimated using the reduced-rank version \( S_y \):
\[
S_y = \begin{bmatrix}
  \Sigma^{-1} & 0 \\
  0 & 0
\end{bmatrix},
\]
(25)
\[
W = V S_y U^T Y.
\]
(26)

### 4. Simulation Study

In the simulation study, the transfer function of the reference model is given in the form of a standard second-order system
\[
G_m(s) = \frac{\omega_m^2}{s^2 + 2\zeta_m \omega_m s + \omega_m^2}
\]
(27)
where \( \omega_m = 20.0 \) and \( \zeta_m = 1.0 \). To obtain the second-order difference equation of the reference model in discrete form as in (6), we used the zero-order hold method [8]. The assumed sampling interval was 20[msec]. Then, the coefficients of the reference model were \( a_m = -1.341, a_{m2} = 4.493 \times 10^{-1}, b_{m0} = 6.155 \times 10^{-2}, \) and \( b_{m1} = 4.714 \times 10^{-2} \), respectively. To study the immunity from the disturbance at \( k = 1500 \) with the magnitude \( d(1500) = 0.25 \) is intentionally added to the plant input \( u(k) \) was also considered. In the simulation, the number of centroids \( N_h \) was arbitrarily fixed to 20, and the number of the tuples \( N \) was chosen as 1000, the value of which is determined by a rule-of-thumb, considering both the nonlinearity and of thereby the ambiguity in the order of the model. For the selection of both \( \sigma \) in (20) and the normalization factor \( \delta \) in (21), it was empirically found that the setting of \( \sigma = 10 \) and \( \delta = N_h \) yields a reasonable performance.

### 4.1. Example 1

In Example 1, the following difference equation for the nonlinear plant with time delay was considered:
\[
y_p(k + 1) = -a_1 y_p(k) - a_2 y_p(k - 1) - \frac{y_p(k)}{1 + y_p^2(k)}
\]
where the coefficients $a_1$, $a_2$, $b_0$, $b_1$, and $L$ were respectively chosen as $a_1 = -1.980$, $a_2 = 9.802 \times 10^{-1}$, $b_0 = 1.987 \times 10^{-4}$, $b_1 = 1.974 \times 10^{-4}$, and $L = 2$. Figure 2 shows a simulation result for the nonlinear plant with time delay with the load disturbance $d(1500) = 0.25$.

4.2. Example 2

In this example, the following difference equation was considered:

$$y_p(k+1) = -a_1 y_p(k) - a_2 y_p(k-1) - y_p^3(k) + b_0 u(k-L) + b_1 u(k-1-L)$$  \hspace{1cm} (29)

with the same coefficients $a_1$, $b_1$, and $L$ as those in Example 1. The simulation result is shown in Figure 3.

Simulation examples are limited to the second-order plants in order to simplify the problem. Notice that a similar design procedure can be easily developed to the higher-order plants. The proposed method uses the RBF-NNs to approximate arbitrary nonlinear mappings. Although the nonlinear mapping can be also performed by the conventional MLP-NNs, we use RBF-NNs because of their stable learning capability and fast training. These networks are widely used to provide the nonlinear controllers as described early.

5. Conclusions

In this paper, a novel design scheme of model-following controllers for the nonlinear plants with time delay has been proposed. It has been shown that the feedback part of the difference equation can be estimated by the RBF-NNs. In the simulation study, it has been demonstrated that the proposed method is not affected by the load disturbance, and shown that the proposed method yields a stable tracking performance. The proposed design scheme is extremely simple as compared with Narendra’s method.

Acknowledgments

This research was supported in part by a research grant from SMC Co., Ltd.

References

Lossless Image Reconstruction by Contractive Mapping of Neighboring Pixels

Hiramatsu Akira†, Korakot Prachumrak‡, Fuchida Takayasu† and Murashima Sadayuki‡

† Dep. Of Information and Computer Science, Fac. of Engineering, Kagoshima University
Koorimoto 1-21-40, Kagoshima 890-0065, Japan
‡Dep. Mathematics & Computer Science, Fac. of Science, King Mongkut’s Institute of Technology
Ladkrabang, Bangkok 10150, Thailand
Email: sc097041@ics.kagoshima-u.ac.jp, mura@ics.kagoshima-u.ac.jp

Abstract—In this paper a method to reconstruct the same image as the original one by copying domain images onto ranges is presented. However this does not mean that an image has self-similarity everywhere. The difference between the calculated image and the original image is kept lower than discrete error on each pixel. The same image as the original one is reconstructed by rounding off of image calculated from the contractive mapping of domain images onto ranges.

1. Introduction

Fractal data compression has been studied for over a decade. At the beginning of its history, the attractive feature of fractal data compression was its high compression ratio. However, expectations for the fractal data compression seemed to diminish, partly, because of the time consuming search for the best-fit domain and also because of the quality of reconstructed image.

Fractal image compression is based on the self-similarity of an image. A fractal image is an image that has self-similarity everywhere. However, in a natural image, self-similarity does not exist everywhere. This is the reason why the scientists do not consider fractal data compression of natural images in the lossless way. A digital image is an image that is constructed on computer screens. The smallest detail on a digital image is the pixel size. The theory in this research is based on this fact. In this paper, a method to reconstruct the same image as the original one by copying domain images onto ranges is presented. The image constructed from affine transformations is referred to as a fractal representation of images. In order to approach this goal, choosing the range as small as possible, i.e., a single pixel range, and the position of the corresponding domain is fixed near the range. The contractive ratio of affine transformation is fixed to 1/9 on size and is fixed to 1/2 on brightness.

The half image plays an important role in this study. The brightness of this image is almost half of the original one. The difference between the half image and the original one is coded as the offset of affine transformation. Here, the term lossless means that the calculated brightness from the affine transformation becomes equal to the original one when the brightness is rounded off. The difference between the initial image and the original one converges to zero in the process of iteration. In next section the basis of the research is presented based on an iterated linear equation system which has the solution equal to the original image. The principles of the lossless representation of digital image are given and verified for the case when the range is a single pixel and the domain consists of 9 pixels for simplicity. It is also shown that the reconstructed image converges to the original one.

2. The Framework of Lossless Fractal Compression of Digital Images

This section presents the concept behind the method of lossless fractal compression of digital images.

2.1. Principles

Lossless fractal image compression method is based on the following principles:
1. Here, an image means a digital image, which is constructed on the screen pixels. An image has no detail finer than a pixel size.
2. This method utilizes the contractive mapping. From the basis of the lossless fractal compression of digital images, an iteration system on the brightness on pixels of the screen has the original image as its solution or as an attractor of the iterated system. The copying of domains onto ranges corresponds to recalculation of all pixel values from their previous values according to the iteration system.
3. The image reconstructed by copying domains onto ranges converges to the original image because the ratio of each contractive mapping on the brightness is set to 1/2. The difference between the initial brightness and the original brightness is multiplied by 1/2 in each iteration and finally diminishes.
4. In order to reduce the data quantity of affine transformations, a kind of discretization is necessary. If the error of the discretization is sufficiently small, the accumulated error can be reduced in the iteration process to be less than 1. It is possible to make the rounded value of brightness to be equal to the original image.
2.2. Screen, Range and Domain

Let \( G \) be a digitized grayscale image consisting of \( M \times N \) pixels. Each pixel's brightness value \( g_{ij} \) can be one of the values 0-255. The reconstructed image from the coding result of \( G \) by the method is \( V \). At fundamental step, the range of an image is set to be one pixel size.

\[
G = \{ g_{ij} | 0 \leq i \leq M - 1, \ 0 \leq j \leq N - 1 \}, \quad (1)
\]

\[
V = \{ v_{ij} | 0 \leq i \leq M - 1, \ 0 \leq j \leq N - 1 \}. \quad (2)
\]

A range consists of a single pixel. Corresponding to the range, the domain is set to be 3x3 pixels near the range as shown in Fig.1.

A range’s brightness value is \( r_0 \). Consequently, a domain includes 9 pixels. The sum of the brightness value of the 9 pixels is represented by \( d_0 \). All ranges cover the whole image without overlapping. The total number of ranges is \( MN \) which are equal to the size of original image. Then,

\[
r_0 = g_{ij}. \quad (3)
\]

The range value \( g_{ij} \) corresponds to 9 pixels of the domain which are \( g_{i-1,j-1}, g_{i-1,j}, g_{i-1,j+1}, g_{i,j-1}, g_{i,j}, g_{i,j+1}, g_{i+1,j-1}, g_{i+1,j}, g_{i+1,j+1} \). By choosing a domain which is close to the range, there is no time consuming search for the best matched domain. Also there is no need to store the information of the domain positions.

2.3. Half Image

A half image plays an important role in this study. It is determined as follows:

\[
G^H = \{ g^H_{ij} | 0 \leq i \leq M - 1, \ 0 \leq j \leq N - 1 \}, \quad (4)
\]

\[
g^H_{ij} = (g_{i-1,j-1} + g_{i-1,j} + g_{i-1,j+1} + g_{i,j-1} + g_{i,j} + g_{i,j+1} + g_{i+1,j-1} + g_{i+1,j} + g_{i+1,j+1})/18. \quad (5)
\]

Fig.1 Single pixel range \( g_{ij} \) and its domain

At the boundary of the image, the cyclic boundary conditions are used as follow:

\[
g_{Mj} = g_{0j}, \quad (0 \leq j \leq N - 1). \quad (6)
\]

The range \( g_{ij} \) is an integer, but the value of \( g^H_{ij} \) is not an integer. The brightness of the half image is nearly half of the original image’s brightness. A half image is a collection of ranges made by the contractive mapping with the ratio 1/9 on size and 1/2 on brightness. The original Mouse (a) and its half image (b) are shown in Fig.2.

(a) (b)

Fig2 “Mouse” and its half image

2.4. A Simultaneous Equation which has the Original Image as Its Solution

Denote \( x_{ij} \) as an unknown pixel value of an image. The simultaneous equation can be written as follows:

\[
x_{ij} = \left( x_{i-1,j-1} + x_{i-1,j} + x_{i-1,j+1} + x_{i,j-1} + x_{i,j} + x_{i,j+1} + x_{i+1,j-1} + x_{i+1,j} + x_{i+1,j+1} \right) / 18 + b_{ij}, \quad (8)
\]

where \( 0 \leq i \leq M - 1 \) and \( 0 \leq j \leq N - 1 \) and the similar cyclic boundary conditions to eqs. (6) and (7) are applied. The value \( b_{ij} \) is determined as follows:

\[
b_{ij} = g_{ij} - g^H_{ij}. \quad (9)
\]

The simultaneous equation is applied in order to construct the original image as \( k \) increases;

\[
x^{(k)}_{ij} \rightarrow g_{ij}, \quad (0 \leq k \leq M - 1, 0 \leq j \leq N - 1). \quad (10)
\]

The simultaneous eq. (8) has a suitable form for an iterated system. At the \( k \)-th iteration, get \( x^{(k)}_{ij} \), so the next calculated value is:

\[
x^{(k+1)}_{ij} = (x^{(k)}_{i-1,j-1} + x^{(k)}_{i-1,j} + x^{(k)}_{i-1,j+1} + x^{(k)}_{i,j-1} + x^{(k)}_{i,j} + x^{(k)}_{i,j+1} + x^{(k)}_{i+1,j-1} + x^{(k)}_{i+1,j} + x^{(k)}_{i+1,j+1}) / 18 + b_{ij}. \quad (10)
\]

The system also converges to the original image, i.e.,

\[
limit_{k \to \infty} x^{(k)}_{ij} = g_{ij} \quad \text{so that the attractor of this iterated system is the original image.}
\]

After repeating the iteration sufficiently, the lossless representation can be obtained by converting \( x^{(k)}_{ij} \) as follows:

\[
v_{ij} = \left[ x^{(k)}_{ij} + 0.5 \right] \quad (11)
\]

\[
g_{iN} = g_{i0}, \quad (0 \leq i \leq M - 1). \quad (7)
\]
3. Coding and Decoding

3.1. Discretization of $b_y$

The iteration system eq. (10) can be implemented to reconstruct the original image. However, 11 bit are needed to represent $b_y$, so it is not suitable for the lossless representation of digital images. Instead of the constant $b_y$, a parameter $m_y$ is introduced. This parameter is given as follows:

$$m_y = g_y - \left[ g_y \right]_{0.5} = \left[ b_y \right]_{0.5}$$ (12)

where $\left[ x \right]_{0.5} = \frac{2x}{2}$. And then,

$$b_y \leq m_y \leq b_y + \frac{8}{18}.$$ (13)

Note that the Discretization error in this process is at most $8/18$. $m_y$ is difference between the original image and the half image.

The decoding process is the iteration of copying domains onto ranges. Let’s rewrite the simultaneous equation by using $m_y$ instead of $b_y$ as follows:

$$x_y^{(k+1)} = (x_y^{(k)} + x_y^{(k)} + x_{i+1,j-1}^{(k)})$$

$$+ x_{i-1,j}^{(k)} + x_y^{(k)} + x_{i+1,j}^{(k)}$$

$$+ x_{i-1,j+1}^{(k)} + x_y^{(k)} + x_{i+1,j+1}^{(k)})/18 + m_y.$$ (14)

In order to consider the convergence characteristics of above iteration process, error between the original image and the calculated one $e_y^{(k)}$ is introduced as follows:

$$e_y^{(k)} = x_y^{(k)} - g_y.$$ (15)

In general,

$$e_y^{(k+1)} = e_y^{(k)} + e_{i-1,j-1}^{(k)} + e_{i+1,j-1}^{(k)}$$

$$+ e_{i-1,j}^{(k)} + e_y^{(k)} + e_{i+1,j}^{(k)}$$

$$+ e_{i-1,j+1}^{(k)} + e_y^{(k)} + e_{i+1,j+1}^{(k)})/18$$

$$+ e',$$

where $e'$ is the discrete error comes from the rounding off in eq. (12) and is given as follows;

$$e' = m_y - b_y,$$ (17)

and as shown in eq. (13) $e'$ vary from 0 to 8/18.

This eq. (16) means that error between the original image and the calculated one is almost halved in every iteration. If continuing this iteration without rounding off, the maximum added error in the second iteration is $8/36$. The maximum added error in the third iteration is $8/72$ and so on. The total maximum error accumulated in this process is $8/18(1+1/2+1/4+...+ 1/2^{k-1})$. Then error between original image and the calculated one $e_y^{(k)}$ is given as follows;

$$\frac{1}{2^k} e_y^{(0)} \leq e_y^{(k)} \leq \frac{8}{18} \left[ 1 - \left( \frac{1}{2} \right)^k \right].$$ (18)

This means that the error of initial value $e_y^{(0)}$ diminishes very rapidly as $k$ increases.

Hence, maximum value of $e_y^{(k)}$ is 8/9 and its minimum value is 0.

Therefore, when $k$ is sufficiently large enough, the following equation can be obtained:

$$g_y \leq x_y^{(k)} \leq g_y + \frac{8}{9}.$$ (19)

After all, calculated value $x_y^{(k)}$ is different from original value $g_y$ by 8/9 in maximum, but the difference is small enough for the reconstructed image to coincide with the original one when rounded off.

The maximum and minimum values of $x_y^{(k)} - g_y$ on “Mouse” converge to the narrow area between 0 and 8/9 as shown in Fig.3.

Consider the above relation of eq. (19), the decoding formula must be as follows:

$$v_y = \left[ x_y^{(k)} + 1/18 \right].$$ (20)
3.2. Data Compression

The amount of $ijm$ is as much as the amount of $ijg$, so we need some more operation to reduce the size of data. Then we formulate $M_{ij}$ from $m_{ij}$ as follows:

$$M_{ij} = m_{ij}/0.5.$$ (21)

After that, $l_{ij}$ is made by applying predictive coding to $M_{ij}$ as follows:

$$l_{ij} = \left(\frac{M_{i,j-2} + M_{i+2,j-2}}{2}\right) - M_{ij}.$$ (22)

The data permuted $l_{ij}$ by this method concentrate on values nearby 0 (Fig.4).

Compression is finished by applying the arithmetic coding to $l_{ij}$. In Tab.2, the result of compression is shown. In this table, compression ratio is defined in eq. (23)

$$C.R. = \frac{\text{Output}}{\text{Input}} \times 100$$ (23)

Table 2 Compression ratio

<table>
<thead>
<tr>
<th>Digital image</th>
<th>Size MxN</th>
<th>C.R. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td>320x200</td>
<td>53.14</td>
</tr>
<tr>
<td>Rose</td>
<td>320x200</td>
<td>62.26</td>
</tr>
<tr>
<td>Lisaw</td>
<td>320x200</td>
<td>55.64</td>
</tr>
<tr>
<td>Lenna</td>
<td>256x256</td>
<td>72.38</td>
</tr>
</tbody>
</table>

3.3. Conclusion

A lossless fractal compression of digital images, in other words, a method to compress a digital image based on contractive mapping of neighboring domains has been presented. Several experiments have been done to show that the original digital image can be accurately reconstructed.

Appendix

Other digital images used in this paper are shown as follows.

Fig.5 “Rose” (320x200)

Fig.6 “Lisaw” (320x200)

Fig.7 “Lenna” (256x256)

References


1058
Abstract – The paper presents a new method for reconstruction of cracks in Nondestructive Electromagnetic Testing (NDET) technology. This is a hybrid method using both stochastic and deterministic optimization techniques. The proposed method combines the advantages of these optimization techniques – global search and reduced number of the objective function evaluations. The hybrid method is implemented as a distributed system of software agents.

1. Introduction

The inverse problem in nondestructive electromagnetic testing consists in the reconstruction of the positions and shapes of cracks in a conducting plate, starting from the signal measured by an Eddy Current Testing (ECT) probe which scans the surface of the plate.

Two different approaches to the solution of the inverse problem are common: the phenomenological approach and the non-phenomenological approach, also known as the model free approach.

The phenomenological approach is based on the comparison between the ECT signal associated with the solution to the electromagnetic field problem corresponding to a proposed (potential) crack and the actual ECT signal. The deterministic methods aiming at the minimization of the distance between the simulated signal and the measured signal (steepest descent, conjugate gradients or Newton) often fail due to the presence of multiple local extremes of the objective function. As well, this approach requires a significant computational effort, is ill conditioned and noise-sensitive, [1].

In the last few years, efficient stochastic heuristic methods (simulated annealing, tabu search, genetic algorithms or evolution strategies) are imposed as standard techniques in complex global optimization problems, such as the inverse problem in nondestructive electromagnetic testing. The main advantages of stochastic methods are their robustness and their ability of finding a global minimum for a known objective function, without requiring the knowledge of the derivative of the objective function, [2]. Unfortunately, these methods require a large number of objective function evaluations and, in the phenomenological approach, each evaluation requires the solving of an electromagnetic field forward problem.

The global solving time is therefore proportional to the CPU (Central Processing Unit) time required to solve the forward problem, multiplied by the number of objective function evaluations needed to find the minimum. In order to reduce the solving time of the forward problem, special computational techniques, called “data-base approaches” were developed, based on finite element, finite element–boundary element and volume integration methods, [3]. Another way to increase the computational efficiency is to reduce the number of function evaluations, in relation with the optimal control of the parameters of the stochastic algorithm (population size or probabilities of genetic operators), [4].

The present paper proposes a combination of stochastic and deterministic optimization algorithms as a new, hybrid method aiming at the reduction of the solving time of the inverse problem in nondestructive electromagnetic testing. Implemented as a distributed system of software agents, this method takes advantage from both the global search achieved by stochastic methods, and the reduced number of the objective function evaluations associated with deterministic optimization methods.

2. Problem Formulation

2.1. The Forward Problem

The structure under study has a three dimensional configuration (figure 1). The conducting plate is characterized by \( \mu = 1 \) and \( \sigma = 30.6 \times 10^6 \) S/m, and has the dimensions \( l_x \times l_y \times h \) (31.5 mm \( \times \) 12 mm \( \times \) 12.5 mm) *.

The plate includes a thin crack of insignificant thickness, located in a suspected region \( a \times b \times c \), and placed in planes of \( y = \text{const} \) (from \( y_1 \) to \( y_2 \) with stepsize \( \text{b/ny} \)). The crack is formed of \((x,z)\) adjacent elementary rectangles, with width \( a/\text{nx} \) and height \( c/\text{nz} \), starting from the scanning surface \((hz = 0 – \text{inner crack})\) or from the hidden surface \((hz = 1 – \text{outer crack})\) of the plate.

The crack is defined by the parameters \( h = [hx_1, hx_2, \ldots, hx_{n_1}, hy, hz]; \)

\( hx_k, \ k = 1, nx, \) is the crack depth corresponding to the \( k \) segment from the \( x \)-axis \((hx_k = 0, \) if there is no crack on the \( k \) segment or \( hx_k \in \{1, nz\} \) if there is crack on the \( k \) segment), \( hy \) is the coordinate \( y \) for the \( z \) plane where is located the crack, while \( hz \) indicates

* The problem TEAM Workshop 15 (Testing Electromagnetic Analysis Methods)
the crack placement (0/1 for inner/outer crack). Figure 1 shows an inner crack with parameters $h = [0, 1, 2, 1, 0, 0, 3, 0]$.

Figure 1: Crack parameterization

The probe, which includes an excitation coil and a measurement coil, is sequentially placed in the center of the $nx \times ny$ rectangles from the $xy$ scanning plane. The evaluated quantity in each scanning point is the variation of the measurement coil impedance between the flawed and unflawed cases.

The impedance variation in the $nx \times ny$ scanning points for a crack with parameters $h$ is obtained solving an electromagnetic problem (the forward problem). In the present paper the forward problem is solved using a volume integration method, [3].

2.2. The Optimization Problem

An optimization problem is formulated as follows: let $f : R^m \rightarrow R$ be the continuous real objective function; the goal is to find $x_0 \in R^m$ such that $f(x_0) = \min_{x \in R^m} (f(x))$.

The function associated with the forward problem is $fd : R^m \rightarrow R^n$, where $m$ is the number of parameters defining the crack, and $n$ is number of parameters defining the information provided by the probe at the scanning points. In the 2-D scanning case under consideration $fd : R^8 \rightarrow R^{72}$ – the segment depths ($hx_k$, $k = 1, 6$), the plane coordinate ($hy$) and the placement ($hz$) define a crack, and the output vector contains the impedance variation (real and imaginary values) in 36 scanning points (the suspected region is discretized in 216 parallelepipedic elements, $nx = 6$, $ny = 6$, $nz = 6$).

If $p^c \in R^m$ is the vector containing the crack parameters to be found, and $v^a \in R^n$ is the information vector provided by the transducer, then the inverse problem can be formulated as an optimization (minimization) problem, with the objective function $f(x) = f_N(fd(x) - v^a)$, where $f_N$ is a norm function.

3. The Hybrid System

The proposed hybrid method, [5], uses both stochastic and deterministic algorithms to solve the optimization problem, with a view to combine the advantages of the two optimization techniques: global search for stochastic methods and reduced number of objective function evaluations for deterministic methods. In order to reach this goal, the stochastic and deterministic optimization programs need to exchange information, according to one of two usual communication models: client/server and point to point.

An agent is an entity which solves the optimization problem and is described by an optimization algorithm, a local memory, a business logic module, and a communication module (Figure 2). The optimization algorithm is used to solve locally the optimization problem, the local memory is used to store data and can not be accessed by other agents, and the communication module is used to interact with other agents, its main functions being the connections management and information exchange. The most important component of the agent is the business logic module; it establishes the optimization method to be used, the interaction moments with other agents, the type and the quantity of information to be exchanged.

Figure 2: The agent and the communication system in the client/server model

A shared memory, a communication module and a business logic module (Figure 2) describe the communication system. The communication module has similar functions with the communication module at the agent level. The shared memory is used to store data and
can be accessed by any agent. The business logic module has three main functions: the management of the shared memory (allocation/deallocation mechanisms), the synchronization for interagent communication (for synchronous systems), and the control of memory data consistency (all the operations executed at the shared memory level have to be atomic).

The hybrid system proposed in this paper for solving the inverse problem of nondestructive electromagnetic testing is an asynchronous distributed system with client/server communication. The system consists in two agents: one agent uses a stochastic algorithm and the other agent uses a free gradient deterministic algorithm. The small number of agents does not influence the system reliability, and the interagent communication overhead is insignificant comparing with one evaluation of the objective function. The main advantages of the client/server communication are scalability, simple parallelization and synchronization mechanism.

The stochastic agent uses a binary genetic algorithm to solve the inverse problem. During functioning it executes only once the iterations/generations stream of the genetic algorithm. After an imposed number of generations of the genetic algorithm, the stochastic agent writes in the shared memory of the communication system the solution with the best fitness measure from the algorithm population. Afterwards, the stochastic agent tries to extract from the shared memory a solution provided by the deterministic agent – the one with the best fitness measure – and incorporates it in the genetic algorithm population if such a solution exists in the shared memory. If the shared memory does not contain a deterministic solution, the communication system notifies the stochastic agent, which resumes the execution without waiting for a solution from the deterministic agent. The communication period (the time elapsed between two consecutive communications) is determined by the number of generations continuously executed by the stochastic agent, and is equivalent to an imposed number of the objective function evaluations. The optimization process stops when the stochastic agent has executed a maximum number of generations (greater than the number of generations for communication).

The deterministic agent uses a free gradient deterministic method to solve the inverse problem. During functioning the deterministic agent executes several times the corresponding algorithm, each time with different initial conditions (solutions). The deterministic agent extracts form the shared memory the solution with the best fitness measure provided by the stochastic agent and uses it as an initialization for the optimization algorithm; then the deterministic agent executes the optimization algorithm and if a better solution is obtained this solution is added in the shared memory. If the shared memory does not contain a stochastic solution, then the deterministic agent waits until the stochastic agent executes an inserting operation.

The communication system controls the exchange of solutions between the stochastic agent and the deterministic agent by providing four functions: insertion/extraction of solutions for the stochastic agent and insertion/extraction of solutions for the deterministic agent. The solutions are stored in the shared memory and in order to maintain the data consistency the insertion/extraction operations have to be atomic: only one agent can access the shared memory at a given time. If an agent tries to execute an insertion operation and there is no space left in the shared memory, then the solution with the worst fitness measure is replaced by the solution to be added. The communication system also provides a stop function, which is called by the stochastic agent to end the optimization process.

4. Results

In order to prove the efficiency of the proposed method two different schemes were used to solve the inverse problem: a simple scheme with one agent using a genetic algorithm and a hybrid scheme with two agents using a genetic algorithm and a deterministic algorithm.

Six inner cracks and three outer cracks were used for testing: The Chebyshev norm was preferred to the Euclidian norm in the objective function expression. In order to make a relevant statistical study each test was repeated for 30 times, each time the genetic algorithm starting from a different random initialization of the population.

In the case of the simple scheme with one agent the genetic algorithm uses a Gray coding and two genetic operators: mutation and one point crossover, with the probabilities 0.1 and 0.8, respectively. The influence of the population size over the performance was studied in this case, the tested values for the population size being 2, 5, 10, 20, 40, 60, 80, 100. The stop criterion is met when a maximum number of generations is reached (equivalent to a maximum number of evaluations of the objective function). The maximum number of evaluations for the objective function was chosen 120000 (under 1% of the searching space ~ 107).

![Figure 3: Optimal population size](image-url)
For each test was determined the number of evaluations such that the objective function to decrease under a value corresponding to the smallest signal of an observable crack.

The best performances were obtained when the population has 40 or 60 individuals (figure 3).

The hybrid scheme was implemented as an asynchronous distributed system with client/server communication. The genetic agent uses an algorithm with the same parameters as the one used in the simple scheme, with a population of 40 individuals (the optimal population size previously found). Two cases were studied for the deterministic agent: in the first case a Simplex algorithm is used, while in the second case a Powell algorithm is used. In both cases the optimal communication period is between 2 and 5. For this interval the performances are better (comparing with the 10–20 interval) with 67% in the Simplex case (for 6 cracks – figure 4) to 77% in the Powell case (for 6 cracks – figure 5).

The hybrid system performs significantly better than the simple system (figure 6): the relative decrease in the number of evaluations of the objective function needed to solve the inverse problem is between 20% and 90%.

5. Conclusions

In the paper is presented a new hybrid method for reconstruction of cracks in nondestructive electromagnetic testing. The method is implemented as a distributed system of software agents and uses both stochastic and deterministic optimization techniques.

The inverse problem is formulated and solved as an optimization problem. The main advantage of the proposed hybrid system is the reduction of the evaluations number for the objective function associated to the optimization problem.

The proposed scheme proved to be more efficient than a simple scheme based on one genetic algorithm. Additionally the performances were studied for a population of test cases. The best performances were obtained when the hybrid system was using for the deterministic agent a Powell based algorithm.

References

Sleep Mode Investigation in Ad Hoc Networks

A. Bisello, A. Giovanardi, G. Mazzini
DI, University of Ferrara, Italy

Abstract—In this paper we propose a sleep mode control scheme for ad hoc networks. Then, we evaluate how the basic parameters of the sleep procedure affect performance in terms of packet delivery success and energy savings. The investigation evaluates also if the packet loss during sleep state affects packet transmission, reception and forwarding in the same way. This trend can help to design an efficient sleep mode procedure, putting nodes in sleep during the more suitable periods.

I. INTRODUCTION

Ad hoc networks are composed by wireless devices communicating each others without any pre-defined infrastructure. Communication between nodes that are not in visibility is enabled by using intermediate nodes in the network as forwarding agents. Nodes in ad hoc networks are often battery operated and have limited energy. Therefore, energy conservation is key for long lived operation. In this paper we focus on the energy optimization at the Network layer. At this layer the energy research efforts can be grouped into three categories: power control or topology control; energy efficient routing; sleep mode control.

If power control is considered, wireless nodes can control the transmission power level [1], with impact on the nodes that are within the radio range. Thus, power level defines the connectivity structure or the topology of the network. Choosing a low power level may lead to a disconnected network, whereas using a very high power level may lead to the consumption of excessive energy.

Given a topology, the energy efficient routing problem is to find an energy efficient route from the source to the destination. The factors affecting path selection are both the cost of transmitting and receiving packets and the resource available at the intermediate nodes. In general, if the transmit power can be modulated in a suitable way, multiple smaller hops are more energy efficient than a single large hop [1]. This is in contrast to the traditional routing where minimum number of hops is a typical route selection criterion.

The sleep mode control problem has the aim to reduce energy consumption by putting a node in a sleeping state a node which is not sending or receiving data. In fact, the energy consumed by a node during the idle state (in which the node frequently senses or listens the channel) is often the greater contribute to the energy consumption. The sleep mode control problem is to identify nodes that can be temporarily frozen and to put them in sleep mode. The problem is not only related to the identification of the moments in which the traffic is generated by or received by a given node, but also the moments in which a node should forward traffic for other nodes.

In this paper we focus on the last problem (sleep mode control), by investigating, by means of a trivial sleep mode protocol, how much the sleep mechanism can affect the final performance in terms of packet loss and energy savings. In the literature some sleep mode protocols have been proposed: GAF [2], SPAN, [3] and APECA [4]. These protocols select, in a high dense ad hoc network, some coordinator nodes which stay always up. Coordinator nodes stay awake continuously and forward packets as necessary. The contrary non-coordinator nodes remain in the sleep state most of the time and wake up only if any data is destined to them or if they have to send data. Any non-coordinator node is associated with a coordinator node which is responsible for buffering any packet destined for it during sleep. The main requirements to select coordinator can be: connectivity,
fairness, energy efficiency, capacity. Connectivity means that enough coordinators should be present so that every node is in the range of at least one coordinator. Fairness means that the coordinators should be rotated in order to ensure that all nodes share the task of forwarding packets. Energy efficiency means that a minimum number of nodes should be coordinators to save maximum energy. Capacity means that enough coordinators should be chosen to preserve the overall throughput of the network. GAF, SPAN and AFECA propose different techniques to select coordinator nodes: they are mainly based on transitions between coordinator and non-coordinator modes chosen by means of probabilistic criteria. These techniques, anyway, often do not guarantee connectivity, fairness or capacity preservation. Furthermore, they do not consider in the coordinator selection the workload (i.e., the traffic load). If workload is known, one can formulate the problem as a scheduling problem where the goal is to create a schedule such that the sleep time of the nodes is maximized. Note that GAF, SPAN and AFECA are designed without a priori investigate how a very simple sleep mode technique can affect final throughput and energy spent, and how link these merit figures to the system parameters.

In this paper we adopt a fully not-coordinated sleep mode protocol, where nodes are put in sleep mode without take care of their position or of the traffic load and then without consider their possible role in the packet generation/forwarding. We simply select a time of activity and a time of sleeping (with given random distribution), which alternate each others. The aim is to identify in a very simple case how the performance fall in term of capacity and increase in term of energy savings, by also identifying the functionalities that are mainly affected by the forced node sleeping: packet transmission, reception or forwarding. To identify the main functionalities affected by the sleep mode (and how they distribute on a given topology) can help to select in the best way the sleep mode protocol. Performance investigation involves the basic parameters characterizing the sleep mode protocol: time of node activity and sleeping time. To perform it we used the Simple Ad hoc siMulator (SAM) [5] that has been modified to: implement the sleep mode protocol, to correctly collect the energy consumption behavior by considering all possible radio states of the wireless network interfaces and to collect all sleep mode control statistics.

II. SLEEP MODE AND ENERGY EVALUATION

In our scheme all nodes alternate states in which they perform all possible network operations, to sleeping states in which they cannot transmit, receive or forward packets. The activity time, called TIME\_ON, is assumed deterministic and constant. When a node has exceeded this time, it enters in a sleep state. The sleeping time is random and can be selected by following different random distributions: UNIFORM, characterized by the maximum sleep time MAX\_SLEEP; GAUS-SIAN, characterized by mean MEAN\_SLEEP and variance VAR\_SLEEP of the sleep time; POISSONIAN, characterized by the average sleep rate LAMBDA\_SLEEP. A node always concludes the current packet transmission before enter in sleep state. All other packets still present in queue and waiting for the transmission are lost when the node goes to sleep (no mechanisms to froze and restore the packet queues are present). On the other hand, reception does not condition in any way the entrance of a certain node in the sleep state. So, a node can suddenly sleep during a packet reception, without take care to ultimate it. Statistics are collected to evaluate if the sleep mode affects in particular packet transmission, reception or forwarding. To evaluate the energy savings introduced by sleeping, we have changed the SAM energy consumption evaluation, by considering all possible radio states assumed by the wireless network interface: transmitting (TX), receiving (RX), idle (IDLE) and sleep (SLEEP). Energy spent is computed by multiplying the power consumption for the time of permanence in a given state. So, for all possible states, we have:

\[ E_{TX} = \text{time}_{TX} \times TX\_POWER_{cons} \]
\[ E_{RX} = \text{time}_{RX} \times RX\_POWER_{cons} \]
\[ E_{IDLE} = \text{time}_{IDLE} \times IDLE\_POWER_{cons} \]
\[ E_{SLEEP} = \text{time}_{SLEEP} \times SLEEP\_POWER_{cons} \]
Power consumption in all states is varying device by device. Anyway, having selected a given device, $TX_{\text{POWER}}$, $RX_{\text{POWER}}$, $IDLE_{\text{POWER}}$, and $SLEEP_{\text{POWER}}$, are constant values. On the other hand, $TX_{\text{POWER}}$ is depending on the transmit power level used. To take into account this dependency we consider:

$$TX_{\text{POWER}} = TX_{\text{POWER}}_{\text{offset}} + F(TX_{\text{POWER}})$$

where $TX_{\text{POWER}}_{\text{offset}}$, expressed in mW, takes into account the power consumption at the input of the RF power amplifier. $F(TX_{\text{POWER}})$ gives the energy spent by the RF power amplifier, depending on the output transmit power $TX_{\text{POWER}}$. This last contribution is strictly related to the amplifier taken into consideration. In general the manufacturer gives datasheet reporting that curve, or that of the absorbed current, as a function of the output power. We have inserted and interpolated (with an interpolation procedure adapted from the Neville’s algorithm) the curve of the MAX2242 2.4 GHz to 2.5 GHz Linear Power Amplifier [6]. To help the selection of the $RX_{\text{POWER}}$, $IDLE_{\text{POWER}}$, $SLEEP_{\text{POWER}}$ and $TX_{\text{POWER}}_{\text{offset}}$ parameters we have inserted some controls at the start of the simulation. If these parameters do not satisfy some specific variation ranges (established by considering many different wireless network devices), an error is returned and a possible range for these values is suggested. All this procedure allows to correctly account for the energy consumption in all possible radio states. Furthermore, we consider each node having at the start of simulation a given limited energy (i.e., not infinite energy), which is decreased at each packet transmission, reception and during idle and sleep states. In this way a given node can switch off when it has exhausted its energy. This implementation can help to better evaluate the energy performance since the sleep mode protocol can be investigated also by means the number of nodes died during simulation.

The effort to account for a correct energy consumption behavior, has been followed by the effort to estimate the right packet loss during SLEEP. SAM has been improved to estimate the number of packets lost during transmission, reception and forwarding as a consequence of the unexpected SLEEP state occurrence. These parameters can be collected by averaging them on the whole network or can be reported node by node. This allows, given a certain topology, to establish the crucial network activities and the nodes more affected by the relaying actions. This investigation could be useful to introduce some mechanisms able to establish the better moments to put a given node in sleeping. For example, by inserting a simple control based on the “drain rate” characterizing a node (i.e., the network load passing through a node), it is possible to better establish if that node could be easily put in sleep mode or not (by delaying its entering in the sleep state). Otherwise, if a node regularly forwards packet for a given node it could advertise that node to delay the sleep entering (also by using piggy-baking), by avoiding packet loss. In this case, minimal modifications are necessary to our trivial sleep protocol and no geographical information (as in the GAF case) or neighbor density information (as in the AFECA case) are needed.

### III. Numerical Results

We consider the following performance indexes:

- **Success probability** $P_{\text{succ}}$: number of correctly delivered packets divided by the number of packets sent;
- **Tot Energy**: total energy spent during simulation (J);

**Fig. 4.** $P_{\text{succ}}$ (up) and Total Energy (down) as functions of $N_{\text{mob}}$.

**Fig. 5.** Percentage of hosts died during simulation as a function of $\text{TIME}_{\text{ON}}$ (up), $\text{MAX}_{\text{SLEEP}}$ (center) and $N$ (down).
• **Percentage of died hosts**: percentage of hosts having exhausted energy during simulation;  
• **Lost packet number**: number of packets lost during TX, RX or forwarding as a consequence of the sleep mode. 

All performance indexes have been averaged during simulation and on the network, i.e., they represent average global values. At the MAC layer we use IEEE 802.11b [7], at the Network layer we consider OLSR [8], at the Transport layer we refer to UDP. In case of mobility the movement is pseudo-linear. The sleep mode protocols parameters are: TIME\_ON and MAX\_SLEEP (with UNIFORM distribution). The main system parameters are reported in Table I. The results are obtained by varying: TIME\_ON, MAX\_SLEEP, N, N\_mob.

Figs. from 1 to 4 show $P_{\text{suc}}$ and Tot Energy as functions of TIME\_ON, MAX\_SLEEP, N and N\_mob, respectively. Observing Fig. 1 (where we vary TIME\_ON and MAX\_SLEEP, with $N = 20$ and $N\_mob = 0$), we can note that, as expected, by increasing the awake time TIME\_ON, increases the success probability, but also the energy consumed. An intuitive opposite trend is present by increasing the maximum sleep time, MAX\_SLEEP. The graphs of Fig. 1 can be used to establish a first simple trade-off between fall in packet delivery success and energy savings: we can fix the tolerated fall in $P_{\text{suc}}$ and the corresponding energy saving and then select the relative sleeping time parameters. Similar considerations hold regarding Fig. 2, where the same performance indexes are reported by exchanging the role of the two sleeping times. As further observation, we can note how a very little increment in MAX\_SLEEP determines a dramatic fall in the success probability and in the energy consumption, while the TIME\_ON dependency seems more soft. Figs. 3 and 4 show $P_{\text{suc}}$ and Tot Energy as functions of N (with $N\_mob = 0$) and N\_mob (with $N = 20$), having considered many TIME\_ON and MAX\_SLEEP combinations. Also in these cases, a trade off between performance fall and energy gain can be achieved. Note that the general trend is quite independent on the number of mobile hosts. Fig. 5 shows the percentage of hosts died during simulations as a function of TIME\_ON (varying MAX\_SLEEP), MAX\_SLEEP (varying TIME\_ON) and N (for many TIME\_ON and MAX\_SLEEP combinations), respectively. Note that by increasing TIME\_ON, N and by decreasing MAX\_SLEEP, the percentage of hosts having exhausted battery increases, as expected. Even if not reported here for brevity, the percentage of died hosts remains the same (quite 60%) by increasing the number of mobile nodes. Finally, Figs. 6 and 7 report the number of packets to be forwarded and having reached the final destination, lost during SLEEP as functions of N (with $N\_mob = 0$) and N\_mob (with $N = 20$), respectively, for many TIME\_ON and MAX\_SLEEP combinations. We can first note that the higher packet loss is addressed to the forwarding procedure, even if many packets are lost also in the last hop of the global path. On the other hand, even if not reported here for brevity, the number of packet lost during transmission is negligible. Furthermore, the lost packet number is increasing with N and is not affected by the number of mobile nodes, N\_mob. As expected, the number of packets lost during sleep decreases by increasing the awake time and by decreasing the sleeping time. Finally, as explained above, all statistics can be used to implement a suitable procedure, able to select the better periods and mechanisms for the sleep mode control.

**REFERENCES**


IC Implementations of Shinriki’s and Inaba’s Chaotic Circuits

Takuya Hamada†, Munehisa Sekikawa‡∗, Yoshihiko Horio†, and Kazuyuki Aihara∗‡

†Graduate School of Engineering, Tokyo Denki University, 2-2 Kanda-Nishiki-cho, Chiyoda-ku, Tokyo, 101-8457, Japan
‡Aihara Complexity Modelling Project, ERATO, JST, 3-23-5-2F Uehara, Shibuya-ku, Tokyo, 151-0064, Japan
* Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, 153-8505, Japan
Email: 06gmd15@ed.cck.dendai.ac.jp

Abstract—Chaotic phenomena observed in electronic circuits have been investigated extensively. Researchers attempt to use chaos in applications such as chaotic communication, chaotic encryption, and combinatorial optimization problems. In particular, chaotic integrated circuits are essential for practical use of the chaotic systems. In this paper, we describe two IC implementations of the three-dimensional autonomous chaotic oscillators, which were proposed by Shinriki and Inaba. These circuits are designed and fabricated through MOSIS TSMC 0.35 μm CMOS process. Measurement results from the chip are reported.

1. Introduction

Nonperiodic oscillations observed in deterministic systems, i.e., deterministic chaos, have been studied widely in a field of natural science. Chaotic phenomena from electronic circuits have been also investigated. Recently, researchers attempt to apply chaos to real-world problems such as chaotic communication, chaotic encryption, and combinatorial optimization problems [1]–[4]. High-dimensional chaos is preferable in these applications. For example, a secure communication exploits the high-dimensional complexity of the chaotic system. Moreover, a large-size combinatorial optimization problem may be solved through high-dimensional chaotic dynamics. Therefore, a compact chaotic integrated circuit is valuable.

Many chaotic circuits have been proposed [5]–[10]. However, most of the continuous-time chaotic circuits are not suitable for an IC implementation, so that they are realized with discrete elements. In this paper, we propose integrated circuit techniques for the three-dimensional autonomous chaotic oscillators proposed by Shinriki [11] and Inaba [12]. For Shinriki’s chaotic circuit, we use an external inductor. In contrast, we replace inductors in Inaba’s chaotic circuit with gyrators. These circuits are designed and fabricated through MOSIS TSMC 0.35 μm CMOS process. We report measurement results from these integrated circuits.

2. Shinriki’s and Inaba’s chaotic circuits

In this section, we briefly explain Shinriki’s chaotic circuit [11] and Inaba’s chaotic circuit [12]. Figure 1 shows Shinriki’s chaotic circuit. This circuit consists of a pair of diodes, two capacitors, an inductor, and a linear negative conductance. In the figure, \(v_1\) and \(v_2\) express the voltages across the capacitors \(C\) and \(C_0\), respectively, and \(i\) expresses the current that flows into the inductor \(L\). The governing equations of the circuit are:

\[
\begin{align*}
L \frac{di}{dt} &= v_1 \\
C \frac{dv_1}{dt} &= -i - i_d \\
C_0 \frac{dv_2}{dt} &= g v_2 + i_d,
\end{align*}
\]

where \(i_d\) is a sum of the currents of the diodes, which is approximately given by \(i_d = (v_1 - v_2)^3\). Figure 2 shows a chaotic attractor observed in Eq. (1).

Figure 1: Shinriki’s chaotic circuit [11].

Figure 2: A chaotic attractor observed from Shinriki’s chaotic circuit with \(C = 0.00872 \, \mu\text{F}, C_0 = 0.00335 \, \mu\text{F}, L = 300 \, \text{mH},\) and \(g = 47.7 \, \mu\text{S}\).

Figure 3 shows Inaba’s chaotic circuit. This circuit comprises a diode, a capacitor, two inductors, and a linear negative conductance. In the figure, \(v\) gives the voltage across
the capacitor \( C \), while \( i_1 \) and \( i_2 \) are the currents of the inductors \( L_1 \) and \( L_2 \), respectively.

Approximating the diode characteristic by Eq. (2), the governing equations of the circuit in Fig. 3 are expressed by a three-dimensional piecewise-linear ordinary differential equations as Eq. (3). A chaotic attractor shown in Fig. 4 is obtained from Eq. (3).

\[
v_d = \begin{cases} \frac{r_1}{V} & (r_1 \leq V) \\ V & (r_1 > V) \end{cases},
\]

\[
\begin{align*}
\frac{dv}{dt} &= -i_1 - i_2 + g v \\
\frac{d}{dt}i_1 &= v - v_d \\
\frac{d}{dt}i_2 &= v.
\end{align*}
\]

3. Integrated circuits and experimental results

Figure 5 shows a photomicrograph of the prototype chip in which Shinriki’s and Inaba’s chaotic circuit shown below are implemented. First, the circuits in Figs. 1 and 3 are modified for the IC implementations. Then, these circuits are fabricated through MOSIS TSMC 0.35 \( \mu \)m CMOS process. Power supply voltages of the chip are \( V_{DD} = 2.5 \) V and \( V_{SS} = -2.5 \) V.

Figure 6 shows a straightforward integrated circuit implementation of Shinriki’s chaotic circuit. Two-stage op-amps are used for the negative conductance and the buffers. The op-amp circuit is shown in Fig. 7. Table 1 summarizes components in the op-amp circuit. The resistors \( R_1 \) and \( R_2 \) in Fig. 6 are \( p \)-diffusion resistors. We choose \( R_3 \) as a variable parameter. We use an external inductor for \( L \). The \( pnp \)-junction diode is constructed with a couple of \( p \)-diffusion and \( n \)-substrate diffusion.

\[
Z_{in} = \frac{R_1 R_3 R_4}{\frac{R_3}{(\pi C)}}
\]

The circuit parameters of the gyrator for \( L_1 \) in Fig. 9 are \( R_1 = R_3 = R_4 = 10 \) k\( \Omega \), \( R_2 = 19.91 \) k\( \Omega \), and
Figure 7: The two-stage op-amp circuit.

Table 1: The circuit parameters for the two-stage op-amp.

<table>
<thead>
<tr>
<th>Elements</th>
<th>Sizes</th>
<th>Fig. 6</th>
<th>Fig. 9</th>
</tr>
</thead>
<tbody>
<tr>
<td>W/L of M₁</td>
<td>10 µm/2 µm</td>
<td>120 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₂</td>
<td>10 µm/2 µm</td>
<td>120 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₃</td>
<td>8.8 µm/2 µm</td>
<td>2 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₄</td>
<td>8.8 µm/2 µm</td>
<td>2 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₅</td>
<td>3.2 µm/2 µm</td>
<td>8 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₆</td>
<td>120 µm/2 µm</td>
<td>248 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₇</td>
<td>128 µm/2 µm</td>
<td>540 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₈</td>
<td>1.6 µm/2 µm</td>
<td>8 µm/4 µm</td>
<td></td>
</tr>
<tr>
<td>W/L of M₉</td>
<td>1.8 µm/2 µm</td>
<td>N/A</td>
<td></td>
</tr>
<tr>
<td>W/L of M io</td>
<td>1.8 µm/3.6 µm</td>
<td>0.8 µm/14 µm</td>
<td></td>
</tr>
<tr>
<td>Cc</td>
<td>2.4 pF</td>
<td>1.1 pF</td>
<td></td>
</tr>
</tbody>
</table>

C = 6 pF, while those for L₂ are, R₁ = R₃ = R₄ = 10 kΩ, R₂ = 10.24 kΩ, and C = 6 pF. Note that parasitic negative conductances exist in the gyrators. Therefore, we connect the external resistors R₂₁ and R₂₂ in parallel to the gyrators in order to control these negative conductances as shown in Fig. 9. In Fig. 9, pn-junction diodes are realized with the p-diffusion and n-substrate diffusion. The gyrators and the buffer in Fig. 9 use the two-stage op-amps in Fig. 7. The components for the two-stage op-amp are shown in Table 1.

The observed attractors from the chip shown in Fig. 5, when we vary the resistor R₁₁, are shown in Fig. 11. These attractors are reconstructed using the time-delay coordinate. That is, the horizontal axis expresses the voltage across the capacitor, and the vertical axis expresses the voltage with a delay of 0.3 µsec. Through the period-doubling bifurcations cascade, nonperiodic attractor is observed as shown in Fig. 11. From these results, we estimated Jacobian matrix, and calculated the maximum Lyapunov exponent. The maximum Lyapunov exponent for the attractor shown in Fig. 11 (d) is 0.049. Therefore, we can conclude that the nonperiodic attractor observed in Fig. 11 (d) is chaotic.

4. Conclusions

In this paper, we have integrated Shinriki’s and Inaba’s chaotic circuits using MOSIS TSMC 0.35 µm CMOS pro-

Figure 8: The attractors observed from the integrated Shinriki’s chaotic circuit with R₁ = R₂ = 10 kΩ, C = C₀ = 10 pF, and L = 30 mH: (a1) period 1 (R₁ = 115.7 kΩ), (a2) period 1 (R₁ = 110.7 kΩ), (b1) period 2 (R₁ = 101.1 kΩ), (b2) period 2 (R₁ = 103.1 kΩ), (c1) period 4 (R₁ = 90.8 kΩ), (c2) period 4 (R₁ = 99.7 kΩ), (d1) chaos (R₁ = 76.2 kΩ), (d2) chaos (R₁ = 84.7 kΩ), and (e) chaos (R₁ = 68.2 kΩ).
cess. In both circuits, period-doubling bifurcations cascade and chaos were observed. As a future problem, we will fully-integrate these circuits, although some external elements have been used in the current implementations.

Acknowledgments

This work was supported in part by Grant-in-Aid no. 16300072 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

On the Estimation of Asymptotic Stability Regions for Nonlinear Discrete-Time Polynomial Systems

Anis Bacha, Houssem Jerbi and Naceur Benhadj Braïek

Laboratoire d’Etude et de Commande Automatique des Processus (LECAP)
Ecole Polytechnique de Tunisie
BP 743, 2078 La Marsa, Tunisia

Email: naceur.benhadj@ept.rnu.tn, houssem.jerbi@fss.rnu.tn, : bachaanis@voila.fr

Abstract—This work presents a discrete reversing trajectory approach for the implementation of the reversing trajectory method for asymptotic stability region estimation of nonlinear discrete polynomial systems. The proposed technique is based upon a polynomial development of the inverse model of the recurrent polynomial considered system. This algebraic formulation is then applied in a four step algorithm yielding a large and similar domain of stability as obtained with the continuous reversing trajectory technique. Based on some topological considerations the systematic computational algorithm is then applied on the second order polynomial model of Hahn. A satisfactory result is obtained which show the benefits of methods to maximize the region of asymptotic stability.

1. Introduction

The problem of estimating the stability domain of an equilibrium point is of fundamental importance for many disciplines in engineering and sciences [2] and [7]. Two main methods for estimating the stability domain of equilibrium points have been proposed in the literature: Lyapunov and non-Lyapunov methods.

The Zubov’s approach [8] offers a technique for computing the entire Region of Asymptotic Stability (RAS) via the optimal Lyapunov Functions. It has been found that for some continuous second order systems, the use of approximated solutions of the linear partial differential equations to construct the optimal Lyapunov functions may result in conservative estimations [1].

The Trajectory Reversing Method, is a non Lyapunov technique [2],[5],[6],[10],[11] and [12], which provides the estimation of stability boundary through a large number of system trajectories obtained by backward integrating the differential equation of the system continuous model. This approach can be considered as very practice and efficient for the enlargement of an initial domain of stability. However this interesting approach has been developed only for continuous systems and no work has considered the discrete non linear case.

In this work we intend to extend the trajectory reversing method for discrete nonlinear systems. For this goal we propose an approach of a polynomial discrete model inversion. The obtained inverse model is applied to perform a backward iteration from the boundary of a guaranteed initial domain of stability, which leads to a large region of asymptotic stability.

Our contribution is presented in this paper in four sections as follows: after this introduction, the section 2 introduces a description of the studied discrete nonlinear polynomial systems and describes the considered problem. In section 3 a new backward iteration approach of discrete nonlinear polynomial model is exposed. Section 4 is devoted to develop an algorithm which presents the main steps of the discrete reversing trajectory method. In the last section an illustrative example of the well known Hahn model [5],[6] is reported to prove the efficiency of the proposed approach.

2. Reversing Trajectory Method Formulation

We consider the autonomous nonlinear discrete-time system given by:

$$X_{k+1} = F(X_k) \tag{1}$$

Without lose of generality we consider that the Function $F(.)$ is analytical and can then be developed into the following polynomial form using Kronecker product [5]:

$$F(X_k) = F_1.X_k + \sum_{i=2}^{r} F_i.X_k^{[i]} \tag{2}$$

where $X_k = [x_{1k} \ x_{2k} \ ... \ x_{nk}]^T \in \mathbb{R}^n$ is the state vector; $F_i,i = 1,...r$ are matrices of $n \times n^i$ dimensions, $r$ is the truncation order and $X_k^{[i]}$ designates the $i^{th}$ order Kronecker power of the state $X_k$. The system (1-2) will be referred as a polynomial system of order $n$ and degree $r$. Particularly quadratic polynomial systems correspond to $(n=2,r=2)$, cubical polynomial ones correspond to $(n=3, r=3)$ and quadratic+cubic polynomial systems correspond to $(n=2, r=3)$.

In the following, we assume that the linear part of the model (1) is asymptotically stable, which means that $F_1$ is a Schur matrix, i.e. the norm of all its eigenvalues are less...
than one $\|\text{eig} (F_i)\| < 1$, and that the polynomial model (2) ensures the existence and the uniqueness of the solution $X(k,X_0)$ for a given initial state $X_0$ for $k=0$.

The region of asymptotic stability of an asymptotically stable equilibrium point $X_e = 0$, considered as the origin, is defined as the set $D$ of all initial points $X_0$, such that:

$$D = \{ X_0 \in \mathbb{R}^n, \ X(k,X_0) \text{ exists for all } k, \text{ and } \lim_{k \to \infty} X(k,X_0) = 0 \}$$

Let $X_e = 0$ be a locally asymptotically stable equilibrium point. Consider $\Omega(X_e)$ to be its stability region and $\Gamma(X_e)$ its boundary. $\Gamma(X_e)$ is invariant and formed by whole trajectories of the system (1) [8]. We shall assume that $\Omega(X_e)$ is not dense in $\mathbb{R}^n$. This is always the case if (1) has more than one stable equilibrium point. The result of the trajectory reversing method is an $(n-1)$-dimensional set of $\mathbb{R}^n$ which constitutes the boundary of a region denoted $\Gamma(X_e)$.

Since, the trajectory reversing method requires the inversion of the polynomial recurrent equation (1), one has to explicit the relation:

$$X_k = F^{-1}(X_{k+1})$$

A polynomial approach to the characterization of this inverse model is proposed in the next section.

3. Proposed technique of nonlinear discrete model inversion

The main problem faced the trajectory reversing method is the inversion of the polynomial recurrent state equation (1-2). An interesting idea is to try to characterize the inverse model as a polynomial form:

$$X_k = F^{-1}(X_{k+1}) = G(X_{k+1})$$

where the vectorial function $G(.)$ is given by:

$$X_k = \sum_{i=1}^{r} G_i X_k^{[i]}$$

(5)

$G_i, i=1,2,...,r$ are matrices of $\binom{n \times n}{i}$ dimensions to determine.

Let’s compute the Kronecker power of $X_k$ presented in (5). We denote:

$${X_k}^r = \sum_{j=1}^{r} G_j X_k^{[j]}$$

(6)

Then, it can be verified that:

$$G_p^2 = \sum_{j=1}^{p-1} (G_{p-j} \otimes G_j)$$

(7)

When replacing $X_k^{[j]}$ in (2) by its expression given (6), it turns out:

$$X_{k+1} = F_1 G_1 X_{k+1} + \left( \sum_{j=1}^{2} F_j G_j \right) X_{k+1}^{[2]} + \ldots + \left( \sum_{j=1}^{r} F_j G_j \right) X_{k+1}^{[r]}$$

(8)

Hence, it is easy to identify the $G_i$ matrices in (5) by solving the following equation matrix system

$$F_1 G_1 = I_n$$

$$F_1 G_2 + F_2 G_2 = 0$$

$$\ldots$$

$$\sum_{j=1}^{i} F_j G_j = 0$$

(9)

for $i=1$ to $r$ where $r$ is a truncation order assumed to be sufficient to have interesting results.

The solutions of the equations (9) are given by:

$$G_1 = F_1^{-1}$$

$$G_2 = -F_1^{-1} F_2 G_2$$

$$\ldots$$

$$G_r = -F_1^{-1} \left( \sum_{i=2}^{r} F_j G_j^{[r]} \right)$$

(10)

1072
4. Algorithm of the stability domain estimation of nonlinear discrete System

Once the inversion of the recurrent equation (1) is performed as presented in previous section, the trajectory reversing method can be applied following a four steps conceptual algorithm:

- Step 1: Determine all equilibriums of (1), i.e. find all solutions of the set of discrete-time polynomial nonlinear equations \( F(X_k) = 0 \).
- Step 2: Study the asymptotic stability of all equilibriums. The asymptotic stability of \( X_e \) follows from the eigenvalues of the linear part of (1). i.e. the norm of all eigenvalues of \( F_i \) are less than one.
- Step 3: For each local asymptotic stable equilibrium point \( X_e \), find an initial set \( \Omega_0 \) where asymptotic stability is guaranteed. This initial region of asymptotic stability of system (1) can be obtained as a ball centered in the origin and of radius \( R_0 \) solution of the following equation [3]:

\[
\sum_{k=2}^{r} \left\| F_k \right\| c^{k-1} R_0^{k-1} \left( 1 - \frac{\alpha}{c} \right) = 0
\]  

(11)

where \( c \) and \( \alpha \) are positive reals verifying the following inequality:

\[
\left\| F_k \right\| c^{k-1} \leq c \alpha^{k-1} \quad \forall k \geq k_0
\]  

(12)

- Step 4 : For each \( X_0 \in \Gamma_0 \), boundary of the guaranteed region of stability \( \Omega_0 \), perform the backward iteration based on (5) and (10). A practical implementation leads to characterize simple connected area \( \Omega_k, i=1,2,...,r \) such that \( \Omega_{k+1} \supset \Omega_k \). These domains are bounded by closed forms \( \Gamma_i, i=1,2,...,r \) which are included in the searched region of asymptotic stability (RAS).

In order to discretize the polynomial model (13) of Hahn system, we have used the approximation of Newton-Raphson [9]. This yields easily to the following expression.

\[
\begin{align*}
\begin{cases}
x_{1k+1} = x_{1k} + \left( - x_{1k} + 2 x_{1k}^2 x_2 \right) T \\
x_{2k+1} = - x_{2k} - x_{2k} T
\end{cases}
\end{align*}
\]  

(14)

where \( T \) is the sampling period; It is easily verified by choosing \( T=0.05 \), that (14) can be presented by the following polynomial state form:

\[
X_{k+1} = F_1 X_k + F_2 X_k^{[3]}
\]  

(15)

where \( F_1 = \begin{bmatrix} 0.95 & 0 \\ 0 & 0.95 \end{bmatrix} \); \( F_3 = \begin{bmatrix} 0 & 0.1 \\ 0 & 0 \end{bmatrix} \)  

As presented in section 3, the backward iteration of polynomial discrete model (15) yields the following reversing recurrent equation:

\[
X_k = G_1 X_{k+1} + G_2 X_{k+1}^{[2]} + G_3 X_{k+1}^{[3]}
\]  

(16)

By referring to equation (10), it turns out that the matrices \( G_1, G_2 \) and \( G_3 \) are respectively given by:

\[
G_1 = \begin{bmatrix} 1.0526 & 0 \\ 0 & 1.0526 \end{bmatrix}; \quad G_2 = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}
\]

The origin is the only equilibrium point of system (15). The autonomous state matrix \( F_i \) presented in (15) is a Shur matrix. The set defined by the origin centered disc of radius \( R_0=0.33 \) is then an asymptotic stability domain of the considered system, which could be considered as an initial guaranteed stability domain for the reversing trajectory method.

By applying the algorithm of the trajectory reversing method as mentioned in the previous section and using the inverse model (16) different regions of asymptotic stability are obtained for different numbers of backward iterations with \( T=0.05s \). The figure 1 represents the region of stability obtained after 20 iterations, and the region of stability represented in the figure 2 is obtained after 50 iterations.

5. Simulation Example

In this section, a possible use of the proposed approach for obtaining a global region of asymptotic stability is discussed for the well known Hahn model. The first issue we want to illustrate is the presentation of the discrete polynomial Hahn model. The continuous polynomial Hahn model is described by the following equations:

\[
\begin{align*}
\begin{cases}
\dot{x}_1 = -x_1 + 2 x_1^2 x_2 \\
x_2 = -x_2
\end{cases}
\end{align*}
\]  

(13)
When comparing the obtained domain of Figure 2 with the domain presented in [6] for the continuous model, one can obviously remark that the two regions of asymptotic stability are identical. This result shows the validity of the proposed discrete approach of the reversing trajectory method for estimating the asymptotic stability domain of discrete non-linear systems.

6. Conclusion

This paper is aimed at the stability domain estimation of the non-linear polynomial discrete-time systems. In this perspective, we have proposed an approach for the discrete polynomial model inversion. This discrete model inversion is used to implement the discrete algorithm of the trajectory reversing method for the asymptotic stability domain determination of non-linear discrete systems.

The proposed technique has been applied to a non-linear discrete model corresponding to the known continuous Hahn model, and interesting result has been reached since the obtained stability domain of the discrete system is really confused with the stability domain of the initial continuous model.

References

Synthesis and Analysis of Aesthetic Fractal Patterns for Textile Design

Takeshi Sakamoto†, Ichiro Kanaya †‡, and Kosuke Sato†

†Graduate School of Engineering Science, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531 JAPAN
‡PRESTO, Japan Science and Technology Agency
Email: sakamoto, kanaya, sato@sens.sys.es.osaka-u.ac.jp

Abstract—Arts and mathematics have very interesting relations. Golden ratio, Fibonacci series, or fractal patterns are well-known examples showing there are mathematical structures behind its beautifulness. The mathematical notation of the relation seems to be interesting to understand, and create, graphical arts/designs. Our researches aim to clarify sensuousness on graphical arts/designs in a scientific manner. In this paper, the authors focus to analyzing human’s impression mechanism about fractal patterns, which are often appeared in the graphical arts and designs. The authors also aim to model mathematically human’s capability to classify impressions about graphic patterns. For this purpose, the authors introduce a couple of impression-scaling methods which quantify impressions of fractal patterns generated by using L-System. Experimental results show these two methods well describe impressions to the fractal patterns of human beings.

1. Introduction

There is a lot of connection with arts and mathematics. Since ancient times, scientists have been deeply interested in the clarification of sensuousness in a scientific manner.

For example, Pythagoras, who was a great philosopher and mathematician in ancient Greece, recognized the relation between music and mathematics for the first time in history. He discovered the music scale that was called Pythagorean scale, and led the development of the following European music.

In area of figurative arts such as architecture, painting, sculpture and so on, the golden ratio is a representative example of the clarification of sensuousness in a scientific manner. If the ratio of two sides of a rectangle is the golden ratio, the rectangle is called “golden rectangle”. It has been believed that the golden rectangle is the most beautiful rectangle. Actually, the golden ratio and the golden rectangle are implicitly or explicitly used for the pyramid, the Parthenon, Picasso and Hokusai’s works, etc.

In the natural world, the golden ratio appears especially in living beings. The golden ratio or Fibonacci series, which is the approximate expression of the golden ratio, is seen in plant structure, spiral of snail, and body structure of human beings, etc. It is possible that we find beautifulness in the golden ratio since that mathematics exists behind the nature.

Arts and mathematics have very interesting relations in those ways; and it is conceivable that clarifying the relations will lead new artistic or mathematical knowledge.

The authors report the recent progress on their research of mathematical notation of relationship among graphical arts/designs and human’s impression about pattern designs. The authors focus to fractal patterns since the fractal patterns are often appeared in the graphical arts and designs.

2. Aesthetic Classification Modeling

It is common knowledge that human can classify the patterns by impression. In this paper, the authors aim to model the aesthetic classification in mathematic manner, which is human’s capability to classify impression about the patterns. In this chapter, the authors describe what mathematical algorithm might be a good model of the aesthetic classification, and deliver how the mathematic model of the aesthetic classification is made. For this purpose, the authors explain the concept of generator and classifier, which are functions that human has for aesthetic classification (see Figure 1).

The generator is a creator of graphical pattern from a certain impression. One of concrete example of the generator is an activity that designers make their works from images in their brains. The classifier is an extractor of certain impression from the pattern. One of concrete example of the classifier is an activity that consumers choose commercial products according to their preference. Human beings have both generators and classifiers. By this way the authors have to model both generator and classifier for human’s ability of aesthetic classification.
The authors explain how they make the mathematical model of the aesthetic classification. First, the patterns are classified by using a certain mathematical algorithm. If the result of this classification and the result of human’s classification are corresponding, the mathematical algorithm can be said to be a model which has the function as a classifier. Next, the patterns are generated by using the mathematical algorithm. If human can classify the generated patterns similarly, this mathematical algorithm can be said be a model which has the function as a generator.

In this paper, the authors propose methods for mathematically modeling the aesthetic classification on fractal patterns, and show experimentally that the methods are effective as classifier. Though it is also necessary to show that the methods are effective as generator, the authors have not worked on this yet; the author choose L-System as a provisional generator in this paper.

### 3. Synthesis of Fractal Patterns

Fractal patterns can be synthesized by using L-System[1]. L-System was invented in 1968 by Aristed Lindenmayer[2], who was a theoretical biologist. The L-System was originally intended to be an algorithm for describing growth of plants. Simple inputs to the L-System generate some complex output that are kind of fractal.

The authors need an amount of different fractal patterns for examining the difference of human’s impression on the different fractal patterns which have mathematical forms. The authors simply explain the algorithm of L-System as follows.

L-System consists of a starting character string and replacing rules applied to it. L-System replaces the starting character string based on the replacing rules, recursively. In consequence, a complex character string with self-similarity is generated. In addition, each character has a geometrical information (‘F’ means drawing a line segment, ‘+’ and ‘−’ means rotating). Finally, by converting the generated character string into geometrical information, the fractal pattern is generated.

Figure 2 shows the appearance that the character string and the fractal pattern are generated, where the starting character string is ‘F’, the replacing rule is ‘F+FX−FX+F’, the angle on ‘+’ and ‘−’ are 60[deg], and the number of iteration is two.

Figure 3 shows the example of fractal patterns generated by L-System. The impression of the fractal patterns are various.

### 4. Methods for Analyzing Fractal Patterns

Impressions of fractal patterns are various from person to person; thus some metric system is required for scaling the impression of the patterns. The authors hereby introduce a couple of impression-scaling methods, named aesthetic fractal convergence, and aesthetic structural components.

#### 4.1. Aesthetic Fractal Convergence

Fractal dimension is a well-known parameter of quantifying the complexity of fractal structure. The authors expand the idea of the fractal dimension to fractal patterns, and propose aesthetic fractal convergence as a parameter that describes intricacy of fractal patterns generated using L-System.

The authors define the aesthetic fractal convergence as follows. Sum of lengths of drawn line segments in fractal pattern is denoted $s$. The length of side of the circumscribed square on the fractal pattern is denoted $l$. The aesthetic fractal convergence is defined as $C = \frac{\log s}{\log l}$ (see Figure 4).

The fractal pattern with great aesthetic fractal convergence is drawn by overcrowding line segments in a small area. As a result, the impression of the fractal pattern becomes intricate (see Figure 5).
4.2. Aesthetic Structural Components

Impressions of figures are often influenced by the impression of their details that compose whole figures. This effect is especially strong when human sees fractal patterns due to its self-similarity. Thus the authors propose aesthetic structural components as a map of describing the composition of fractal patterns.

To explain the aesthetic structural components, partial figure and partial figure map are defined as follows. The partial figure is defined as a figure that consists of three-line segments of equal lengths. These three-line segments make two angles, $\theta_1$ and $\theta_2$ (see Figure 6). The partial figure map is defined as a map where partial figures are placed with $\theta_1$ and $\theta_2$ as axes (see Figure 7).

By using the partial figure and the partial figure map, aesthetic structural components is obtained as follows. To analyze the components of fractal pattern, the fractal pattern is resolved to partial figures. Next, the ratio in the obtained partial figures is mapped to partial figure map as a gray-scale value (see Figure 8). The authors denote this map the aesthetic structural components.

The bit maps in the lower part of Figure 9 is the corresponding aesthetic structural components of the fractal patterns shown in the upper part of Figure 9. If a couple of aesthetic structural components are similar, the impressions of them are also the closer.

Components similarity, $M$ is defined as the regularized sum of absolute differences (SAD) of the aesthetic structural components,

$$M = 1 - \sum_{i,j} \frac{|a_{ij} - b_{ij}|}{2}$$

where $a_{ij}, b_{ij}$ are $i$-th elements of the aesthetic structural components of two fractal patterns, say, $A$ and $B$, respectively. If $M$ of two comparing figures is the greater, components of those two figures are the closer, and the impressions of them are also the closer.

5. Experiments and Results

The authors verify whether aesthetic fractal convergence and components similarity well describe impressions to fractal patterns of human beings. The fractal patterns targeted by these experiments are limited on the following parameters; the starting character string is ‘F’, the replacing rule has four ‘F’, and the number of iteration is four.

5.1. Experiment and Result on Aesthetic Fractal Convergence

The purpose of this experiment is to verify that there is a correlation between aesthetic fractal convergence $C$ and intricateness that examinees feel in fractal patterns. Examinees were ten persons from 22 to 25 years old. Nine fractal patterns (each patterns are printed separately on pieces of paper) whose $C$ was in the range of $1.0 < C \leq 1.5$ were prepared. They were brought together as group A. Similarly, nine fractal patterns whose $C$ was in the range of $1.5 < C \leq 2.0$ were brought together as group B. In this way, ten groups from A to J were prepared at intervals of 0.5 aesthetic fractal convergence.

The ten sample groups were presented to the examinees. The examinees were to locate the samples (pieces of paper) on a workspace (desktop) labeled from 0[cm] to 70[cm], based on their feelings of intricateness of each samples. The examinees expressed their degree of intricateness of the samples by distance from the edge of the space, where position at 0[cm] (the leftmost position) meant the least
intricateness and position at 70[cm] (the rightmost position) meant the most intricateness. These positions were recorded as subjective intricacy of the sample.

Figure 10 shows the relation between aesthetic fractal convergence and subjective intricacy obtained by this experiment, where the subjective intricacy is an average of all examinées. Correlation coefficient between aesthetic fractal convergence and subjective intricacy was 0.98.

5.2. Experiment and Result on Components Similarity

The purpose of experiment is to verify that there is a correlation between components similarity $M$ and similarity (or likeness) that human feels from fractal patterns.

Examinées were ten persons from 22 to 25 years old. A single fractal pattern was selected as a standard pattern, and other 20 patterns with different components similarity to the standard pattern were prepared as sample patterns. The samples are selected in order to reduce the influence of aesthetic fractal convergence on the impressions, the authors set conditions on this experiment that the difference between standard pattern and each left patterns on fractal convergence didn’t exceed 0.4.

First, the standard pattern is presented to the examinées, and it was put on the edge of the workspace (desktop) labeled 0[cm]. Next, 20 sample patterns were presented to the examinées. The examinées locate the sample patterns between 0[cm] and 70[cm] of workspace, based on similarity by their feelings. These lengths were recorded as subjective similarity of the sample.

Figure 11 shows the relation between components similarity and subjective similarity obtained by this experiment, where the subjective similarity is an average of all testees. As a result, the correlation coefficient between components similarity and subjective similarity was 0.76.

6. Discussion

There is a very strong correlation between subjective intricacy and aesthetic fractal convergence; the correlation coefficient of them is 0.98 from the experimental result on aesthetic fractal convergence. It can be said that the aesthetic fractal convergence is a parameter that expresses well intricateness feelings in fractal patterns. It is conceivable that the impression of fractal patterns can be classified by the proposed aesthetic fractal convergence.

There is a strong correlation between subjective similarity and components similarity; the correlation coefficient of them is 0.76 from the experimental result of components similarity. It can be said that the components similarity is a parameter that expresses well similarity feelings in fractal patterns.

There were, however, some problems remained in this experiment. The proposed components similarity considers only a similarity of partial components in fractal pattern and doesn’t consider what shape the entire pattern is. Human beings feel a couple of figures similar if the silhouette of those figures are alike, even though the component of them are quite different. Some influences of entire shapes figures to the human’s impression are to be considered.

7. Conclusion

This proposal aims to discover how human beings feel their impression from the graphic patterns. In this paper, the authors analyzed impression of fractal patterns generated by using L-System, by introducing a couple of impression-scaling methods named aesthetic fractal convergence and aesthetic structural components. Experimental results show these two methods well describe impressions to the graphic patterns of human beings.

References


Improved Parallel Processing Hardware Algorithm for Large-Scale Quadratic Assignment Problems

Naoki Ogawa †, Yuya Ohashi ‡, Yoshihiko Horio † and Kazuyuki Aihara *

†Graduate School of Engineering, Tokyo Denki University, 2–2 Kanda–Nishiki–cho, Chiyoda–ku, Tokyo, 101–8457, Japan. Email: 05gmd06@ed.cck.dendai.ac.jp Tel: +81–3–5280–3362
‡Aihara Complexity Modeling Project, ERATO, JST, 3–23–5 Uehara, Shibuya–ku, Tokyo, 151–0064, Japan
* Institute of Industrial Science, University of Tokyo, 4–6–1 Komaba, Meguro–ku, Tokyo, 153–8505, Japan

Abstract—The quadratic assignment problem (QAP) is one of the nondeterministic polynomial (NP)-hard combinatorial optimization problems. We have implemented the chaotic exponential tabu search algorithm into a mixed analog/digital hardware system for size-10 QAPs. Furthermore, we have proposed two parallel processing algorithms to improve the speed of the system, at the same time, to extend the system size for a large-scale QAPs. In this paper, we further improve the parallel algorithm, in particular, taking the hardware constrains into account. First, a novel formulation of the QAP with 4-dimensional matrix instead of two 2-dimensional matrices in a standard formulation is proposed. Then, we propose a practical parallel algorithm suitable for a modular mixed analog/digital hardware architecture.

1. Introduction

The quadratic assignment problem (QAP) is one of the NP-hard combinational optimization problems [1]. Obtaining the optimal solution of the QAP requires impractical computational time. Currently, the optimum solution of the QAP with size more than 36 is unknown [1]. Therefore, heuristic methods are very important to find good near optimum solutions in reasonable time.

Tabu search with 2-opt algorithm is one of such heuristic methods [2], which escapes from undesirable local minima by using a tabu list. Once a certain 2-opt exchange is stored in the tabu list, the same exchange is forbidden for x iterations, where x is a tabu list size, and becomes available after x iterations.

Hasegawa et al. proposed an implementation of the tabu search on a neural network [3–5]. Furthermore, they extended the ordinary tabu to an exponential tabu by using exponential decay of the refractoriness of a neuron model. Moreover, they introduced chaotic search into the exponential tabu search by replacing static neurons with chaotic neurons [3–5]. As a result, they confirmed superior performance of the chaotic exponential tabu search in solving the QAP through computer simulations.

We proposed an efficient mixed analog/digital circuit hardware for the chaotic exponential tabu search algorithm [6], [7]. The hardware system uses switched-current chaotic neuron integrated circuits for a rapid implementation of physical chaotic dynamics [8], [9]. However, the hardware system uses a sequential update of a neuronal states, so that n x n updates (n is the size of the QAP) are required for one iteration of the algorithm. Therefore, it would take a large amount of time for a large-size QAP. To overcome this problem, we have proposed two parallel processing algorithms [10].

In this paper, we further improve the parallel algorithms under the hardware constrains. First, we use a 4-dimensional matrix to calculate the objective function of the QAP instead of two 2-dimensional matrices used in a standard formulation. The scaling and quantization of the original problems, which is necessary for a hardware implementation, are done using the 4-dimensional matrix. (We do not mention these in this paper because of the space shortage.) Then, we propose an improved parallel processing algorithm suitable for a modular hardware architecture for a large-scale QAP. Finally, we discuss the number of neurons, which are simultaneously updated in the parallel algorithm, in order to obtain a good balance between the hardware overhead and the processing speed.

2. Quadratic Assignment Problem Formulation with a 4-Dimensional Matrix

The QAP of size n consists of n locations and n units. An n x n “distance” matrix denotes mutual distances among the locations. Moreover, an n x n “flow” matrix expresses mutual relations among the units. The QAP is defined such that we should find an assignment of the units to the locations that minimizes the cost function $FP$ given by

$$FP = \sum_{g=1}^{n} \sum_{h=1}^{n} a_{gh} b_{p(g)p(h)} = \sum_{g=1}^{n} \sum_{h=1}^{n} c_{gh} p(g)p(h).$$  

(1)

where $p$ is a permutation of n elements given by eq. (2). 

which expresses one of the feasible solutions. $a_{gh}$ is the distance between the locations $g$ and $h$, $b_{p(g)p(h)}$ is the flow between the units $p(g)$ and $p(h)$, and $p(g)$ represents the $g$th element of the permutation $p$.

$$p : (p(1), p(2), \ldots, p(g), \ldots, p(h), \ldots, p(n)).$$  

(2)
If the current permutation $p$ gives the minimum of $FP$, $p$ is the optimal solution.

We now introduce a new 4-dimensional matrix $c_{gh(p(p(q))p(h))}$ given by the product of $a_{gh}$ and $b_{p(p(q))p(h)}$. The use of pre-calculated 4-dimensional matrix drastically reduces the processing speed and hardware requirements because the calculation of $FP$, as a result, does not involve any multiplication but only additions as shown in section 4.

### 3. Chaotic Exponential Tabu Search

Assuming the size of the problems is $n$, the neural network composed of $n \times n$ chaotic neurons as shown in Fig. 1(a) is prepared. In [3]–[5], the state of each neuron is updated one by one from the (1, 1)th neuron in the network to the ($n$, $n$)th neuron. We denote this update sequence as one “iteration.” If the (i, j)th neuron in the network fires on the course of updating, the element $i$ of the permutation $p$ is assigned to the index $j$ as shown in Fig. 1(b). At the same time, the element $p(j)$ is assigned to the index $q(i)$. These exchanges are referred to “(i, j)” and “(p(j), q(i))” assignments, respectively.

![Figure 1](image)

**Figure 1:** (a) The neural network for the QAP of size $n$, and (b) the permutation $p$ and the (i, j) and (p(j), q(i))-assignments.

The chaotic dynamics of the (i, j)th neuron in the network used in our hardware system is defined as follows [11]:

$$
\xi_j(t + 1) = \beta[FP(t) - FP_i(t)],
$$

$$
\eta_{ij}(t + 1) = k_f\eta_{ij}(t) - \alpha\xi_{ip}(t) + R,
$$

$$
\zeta_{ij}(t + 1) = k_c\zeta_{ij}(t) - 3\alpha\xi_{ip}(t) + R,
$$

$$
x_{ij}(t + 1) = f[\xi_{ij}(t + 1) + \eta_{ij}(t + 1) + \zeta_{ij}(t + 1)],
$$

where $FP_i(t)$ is the current cost function, $FP_j(t)$ is the cost after the (i, j)-assignment. Therefore, the internal state $\xi_{ij}(t + 1)$ gives the gain of the cost function resulting from the (i, j)-assignment, where $\beta$ is a scaling parameter of the gain. Moreover, the internal state $\eta_{ij}(t)$ is a sum of feedback from other neurons and tabu effect for the (p(j), q(i))-assignment, the internal state $\zeta_{ij}(t)$ is the tabu effect for the (i, j)-assignment, $\alpha$ is a scaling parameter for the tabu effect, $k_f$ and $k_c$ are decay parameters for the tabu effect, $R$ is an external bias, and $f(\cdot)$ is a monotonically increasing nonlinear output function of the neuron. During the sequential updating process, the (i, j)th neuron fires if $x_{ij}(t + 1) \geq 0.5$.

### 4. Gain Calculation Using the 4-Dimensional Matrix

Eq. (3) gives the gain resulting from the (i, j)-assignment. By substituting eq. (1) to eq. (3), the gain can be simply calculated with the 4-dimensional matrix as:

$$
FP_i(t) - FP_j(t) = c_{ijp(p(q))p(j)} - c_{ijq}
$$

$$
+ c_{ijp(p(q))p(j)} - c_{ijp(p(q))p(j)}
$$

$$
+ c_{ijp(p(q))p(j)} - c_{ijp(p(q))p(j)}
$$

$$
+ c_{ijp(p(q))p(j)} - c_{ijp(p(q))p(j)}
$$

$$
+ \sum_{k=1}^{n} (c_{kj}(k)p(j) - c_{kj}(k)p(j))
$$

$$
+ c_{kj}(k)p(k) - c_{kj}(k)p(j)
$$

$$
+ c_{kj}(k)p(k) - c_{kj}(k)p(j)
$$

$$
+ c_{kj}(k)p(k) - c_{kj}(k)p(j),
$$

where $c_{ijkl}$ is the (i, j, k, l)th element of the 4-dimensional matrix defined in eq. (1). It should be underscored that we don’t need any multiplication in eq. (8) thanks to the 4-dimensional matrix.

### 5. A Hardware System Architecture Using Parallel Algorithm

We constructed the prototype hardware system and confirmed good performance of the system [6], [7]. However, the maximal size of the QAP for the prototype system was
In this section, we propose a modular hardware system architecture for large-scale QAPs. In order to improve the speed and efficiency of the system, and at the time, to minimized the hardware overhead, we propose an improved parallel processing algorithm based on the predecessors in [10].

### 5.1. Parallel algorithm

The simultaneous calculation of the gains given by eq. (8) for some neurons, of which internal states are also updated simultaneously, realizes a parallel processing scheme for the exponential chaotic tabu search hardware system. The details are as follows:

We first pick up \( m \) consecutive neurons in the network starting from the first neuron in the 2-dimensional network shown in Fig. 1(a), i.e., the \((1, 1)\)th neuron. Then, we update all of these \( m \) neurons, and calculate the gains of the cost function for each of these neurons simultaneously. Finally, we chose the neuron that has the largest internal state value \( y_{ij} \). If the neuron with the largest \( y_{ij} \) fires, we execute the 2-opt exchange.

If the parallel number \( m \) is changed depending on the problem, the hardware system will be very complex. Therefore, we should fix the number \( m \). With the fixed value of \( m \), the neurons in the network are updated as follows: For example, when we solve the size-17 QAP with \( m = 5 \), a total of \( 17 \times 17 \) neurons are used from the entire network as shown in Fig. 2(a). Then, we make groups of 5 neurons as shown in Fig. 2(b). That is, the first neuron group consists of the \((1, 1)\)th, \((1, 2)\)th, \((1, 3)\)th, \((1, 4)\)th, and \((1, 5)\)th neurons, the second group is composed of the \((1, 6)\)th, \((1, 7)\)th, \((1, 8)\)th, \((1, 9)\)th and \((1, 10)\)th neurons, and so on. Thus, the final neuron group consists of \((17, 14)\)th, \((17, 15)\)th, \((17, 16)\)th, \((17, 17)\)th and \((18, 1)\)th neurons. Finally, each group of neurons serially updated according to the system clock.

The above procedure is suitable for a modular hardware system. In such system, some numbers of the neurons, which is an integer multiple of \( m \), are deployed on a unit circuit board. Moreover, several unit boards compose the whole system.

### 5.2. Neuron number \( m \)

The larger \( m \) is, the less calculation clock cycle, \( t_{\text{total}} \), is as given in eq. (9).

\[
\frac{t_{\text{total}}}{m} = \frac{\text{Total Iteration Number}}{m} \quad (9)
\]

However, if \( m \) is large, the hardware system becomes also large in size. Furthermore, with large \( m \), it will be difficult to decide the largest internal state value among \( m \) neurons, in particular, with an analog circuitry. We will determine the value of \( m \) through simulations in the next section.

![Figure 2: (a) The neural network with 400 neurons, 289 neurons are used for a size-17 QAP. The rest of the neurons are unassigned. (b) When \( m = 5 \), neurons in each of 5-neuron group are updated simultaneously. The group is taken serially according to the system clock.](image-url)

### 6. Simulation Results

We solve the benchmark problems, Tai12a, Tai15a, and Tai17a from the QAP library [1] using the proposed parallel processing algorithm. We define the normalized average clock cycles (NACC) for obtaining the optimal solution as

\[
\text{NACC} = \frac{\text{AIN} / m}{\text{AINWOPP}} \quad (10)
\]

where AIN is the average iteration number to obtain the optimal solution using the parallel processing algorithm, and AINWOPP is that without any parallel scheme, that is, using the original sequential update. The iteration numbers are averaged over 100 trials with different initial conditions.

Table 1 shows the NACC with the parallel algorithm when \( m \) is changed. As shown in the table, the processing
speed is improved as $m$ increases. However, the efficiency of the chaotic search does not depend on $m$. Moreover, the small $m$ would be preferable for the hardware implementation. Furthermore, as shown in the table, the solving speed is much improved than the original algorithm even with $m = 5$. As a consequence, we conclude that we use $m = 5$ in our hardware system.

Table 1: Summary of the simulation results. The NACC given in eq. (10) are shown for different values of $m$.

<table>
<thead>
<tr>
<th>Instances</th>
<th>Tai12a</th>
<th>Tai15a</th>
<th>Tai17a</th>
</tr>
</thead>
<tbody>
<tr>
<td>original</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>$m = 5$</td>
<td>0.26</td>
<td>0.24</td>
<td>0.24</td>
</tr>
<tr>
<td>$m = 10$</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>$m = n$</td>
<td>0.11</td>
<td>0.08</td>
<td>0.07</td>
</tr>
</tbody>
</table>

7. Conclusions

We have improved the parallel algorithm proposed in [10] taking the hardware constrains into account. Moreover, we have modified the chaotic neuron model for the hardware system. A novel formulation of the QAP with the 4-dimensional matrix instead of two 2-dimensional matrices in a standard formulation have been proposed. As a result, the gain calculation became simple enough for an analog circuit implementation. Furthermore, we have proposed a practical parallel algorithm suitable for a modular hardware architecture. Moreover, we have shown that the proposed algorithm gives improvement in speed.

In the future, we will consider a scaling and quantization method of the 4-dimensional matrix in detail. Then, a hardware-efficient nonlinear scaling and quantization scheme will be proposed. Finally, we will physically construct a hardware system for large-scale QAPs.

Acknowledgments

The authors would like to thank T. Ikeguchi of Saitama University and M. Hasegawa of Communications Research Laboratory for their valuable discussions. This work was supported in part by Grant-in-Aid no. 16300072 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References


Special Session Organizer index

A
Aihara Kazuyuki .................................................. 27
Amigo Jose .................................................. 25
Asai Hideki .................................................. 33

G
Gohara Kazutoshi .......................... 10, 12, 15

H
Hikihara Takashi .................................................. 20, 31
Hirata Yoshito .................................................. 14, 17

I
Inuiguchi Masahiro .................................................. 10

K
Kocarev Ljupco .................................................. 25
Kokubu Hiroshi .................................................. 20

L
Lau Francis C.M. .................................................. 22

M
Matsumoto Akio .................................................. 14, 16
Mezic Igor .................................................. 31

N
Nakajima Koji .................................................. 21, 23
Nakamura Tomomichi .................................................. 14, 17
Nakashizuka Makoto .................................................. 26

O
Oishi Shin’ichi. .................................................. 16

P
Plum Michael .................................................. 19

R
Rakoto-Ravalontsalama Naly .................................................. 15, 18
Raman Arvind .................................................. 31
Rump Siegfried M. .................................................. 16

S
Saito Toshimichi .................................................. 10, 12, 15
Schwarz Wolfgang .................................................. 10, 12, 15
Shinoda Shoji .................................................. 20, 23
Small Michael .................................................. 14, 17
Stoop Ruedi .................................................. 33

T
Takai Shigemasa .................................................. 15, 18
Tse Chi K. .................................................. 10, 12, 15, 22
Tsujii Shigeki .................................................. 27
Tsumoto Kunichika .................................................. 27
Tsurumi Masayo .................................................. 10

U
Uchida Atsushi .................................................. 15, 19
Ushio Toshimitsu .................................................. 10, 12, 15

W
Wenck Florian .................................................. 18

Y
Yamamoto Nobito .................................................. 19
This page has been intentionally left blank.
## Chairman index

### A
- Amigo Jose ........................................... 25
- Arrowsmith David ................................... 25
- Asano Akira ........................................... 26

### B
- Bizzarri Federico ..................................... 34
- Brambilla Angelo ..................................... 13

### C
- Callegari Sergio ...................................... 17

### D
- De Feo Oscar .......................................... 11

### E
- Endo Tetsuro .......................................... 13
- Endo Yasunori ........................................ 27

### F
- Fujisaka Hisato ....................................... 22

### G
- Gilli Marco ........................................... 26
- Gohara K. ............................................. 12

### H
- Hasler Martin ......................................... 9
- Hayakawa Yoshihiro ................................... 23
- Hayashi Hatsuo ........................................ 23
- Hikihara Takashi ...................................... 31
- Hirata Yoshito ........................................ 14, 17
- Hisakado Takashi ..................................... 35
- Hosokawa Yasuteru ................................... 11

### I
- Inoue Kaoru ........................................... 21
- Inuiiguchi Masahiro ................................... 10

### J
- Jin’no Kenya ........................................... 24

### K
- Jitsumatsu Yutaka ...................................... 28
- Kanaya Ichiro ......................................... 35
- Kanzawa Yuchi ......................................... 27
- Kennedy Peter ......................................... 24
- Kitajima Hiroyuki ...................................... 34
- Kocarev Ljupco ........................................ 25
- Kokubu Hiroshi .......................................... 20
- Kolumban Geza ......................................... 28
- Komatsu Kazuo ......................................... 30
- Konishi Keiji ........................................... 22
- Kurokawa Hiroyuki ..................................... 16
- Kwak Kyung Sup ........................................ 20

### L
- Lansky Petr ........................................... 24
- Lau Francis ............................................ 22
- Lawrance Antony J. ................................... 25

### M
- Makino Mitsunori ...................................... 28
- Matsueda Hideaki ...................................... 31
- Matsumoto Akio ........................................ 14
- Mazzini Gianluca ....................................... 32
- Miyano Takaya .......................................... 12
- Mizutani Shin .......................................... 29
- Moro Seiichiro ......................................... 11
- Murashima Sadayuki ................................... 34

### N
- Nakagawa Masahiro ................................... 12
- Nakajima Koji ........................................... 21
- Nakamura Tomomichi .................................. 14, 17
- Nakaniishi Takahiro ................................... 11
- Nakano Keisuke ........................................ 23
- Nakashizuka Makoto ................................... 26

### O
- Ogorzalek Maciej ...................................... 31
- Ohsumi Akira ........................................... 29
- Ohta Yuzo ................................................ 30
- Oishi Shin’ichi .......................................... 16

### P

1085
Plum Michael ........................................... 19

R
Rovatti Riccardo ................................. 35
Rump Siegfried M. ................. 16

S
Sato Shigeo ........................................... 21
Schimming Thomas ......................... 29
Schwarz Wolfgang ...................... 10, 21
Setti Gianluca ................................. 9
Shigei Noritaka ......................... 28
Small Michael ......................... 14, 17
Stoop Ruedi ................................. 13, 33
Susuki Yoshihiko ................. 24

T
Takahashi Norikazu ......................... 26
Takai Shigemasa ............................. 18
Tanaka Hisa-Aki ......................... 32
Tanji Yuichi ................................. 33
Tatsumi Keiji ................................. 29
Torikai Hiroyuki ......................... 15
Tse C. K. ................................. 12
Tse Chi K. ................................ 22
Tsubone Tadashi ......................... 13
Tsujii Shigeki ............................. 27
Tsumoto Kunichika ....................... 27
Tsuneda Akio ................................ 17
Tsurumi Masayo ......................... 10

U
Uchida Atsushi ................................. 15, 19
Ushio Toshimitsu ......................... 9, 15, 32

W
Wada Masahiro ................................. 9
Wenck Florian ............................... 18

Y
Yamamoto Nobito ................................. 19
Yoshimura Kazuyuki ....................... 15
Author index

Notice:

- Page numbers using underlined characters point to full papers.
- Page numbers using regular characters point to sessions in the conference program.

A
Acciani Giuseppe ........................................... 33, 991
Addabbo Tommaso ......................................... 17, 25, 375, 663
Aihara Ikkyu .................................................. 27, 767
Aihara Kazuyuki ...................... 17, 21, 27, 34, 35, 363, 495, 499, 747, 751, 755, 759, 767, 1031, 1067, 1079
Al-Naimee Kais ............................................. 19, 447
Alioto Massimo .............................................. 17, 375
Amemiya Yoshihito ......................................... 32, 975
Amigó José M. .............................................. 25, 28, 651, 815
An Xizhi ....................................................... 23, 575
Anis Bacha ..................................................... 35, 1071
Aomori Hisashi ......................... 11, 13, 28, 115, 211, 819, 823
Aono Masashi .................................................. 27, 763
Arai Kenichi .................................................... 29, 843
Arai Kensuke .................................................. 15, 295
Arai Shintaro ................................................... 28, 799
Arai Zin .......................................................... 20, 467
Arcelli Fortunato T. .................. 19, 447
Arrowsmith David K. ................ 25, 675
Arthur Benjamin I ......................... 23, 607
Asada Toichiro ............................................... 14
Asai Hideki ..................................................... 33, 1007
Asai Tetsuya .................................................... 32, 975
Asano Akira ..................................................... 26, 711
Ascoli Alon ...................................................... 22, 547
Atanov Nikolay ............................................... 28, 795
Atarashi Hiroyuki ........................................... 20, 491
Azuma Shun-ichi ............................................. 15, 267

B
Bai Yongming ................................................... 18, 411
Balestra Michele .............................................. 15, 259
Balibrea Francisco ........................................... 17, 371
Banerjee Soumitro ......................... 10, 12, 63, 163
Belghith Safya .................................................. 12, 147
Bellini Alberto .................................................. 21, 511
Belykh Igor ..................................................... 22, 567
Belykh Vladimir ............................................... 22, 567
Bévilacqua Alessandro .................. 26, 29, 715, 831
Bhattacharya Tapas K. .................. 12, 163
Biey Mario ...................................................... 25, 671
Bisello Alberto ............................................... 34, 1063
Bizzarri Federico ............................................. 34, 1039
Blakely Jonathan N. .......................... 34, 1047
Bolle Desire ..................................................... 30, 879
Bonfè Marcello ............................................... 27, 775
Bouzouache Hajer .......................................... 30, 899
Boyd Stephen P. .............................................. 9
Brambilla Angelo ............................................. 9, 39
Buonomo Antonio ........................................... 13, 195
Buscarino Arturo ............................................. 9, 51

C
Caballero M.Victoria ....................................... 17, 371
Cabasino Maria P. .......................................... 18, 399
Cafagna Donato .............................................. 28, 34, 827, 1043
Callegari Sergio ............................................. 18, 387
Canu Michael ................................................... 15, 275
Caporale Salvatore ......................................... 29, 855
Carmeli Cristian .............................................. 14, 247
Casaccia Stefano ............................................. 16, 335
Chan Chi-Keung ............................................ 11, 127
Chargé Pascal .................................................. 28, 811
Checco Paolo ................................................... 25, 671
Chen Guanrong ............................................... 22, 555
Chen Jun-Ning .............................................. 12, 159
Chen Luunan ................................................... 27, 779
Chen Yanfeng ................................................... 12, 175
Chiesi Lorenzo ............................................... 21, 511
Cimatti Giampaolo .......................................... 23, 583
Cincotti Silvano ............................................... 16, 335
Claus Martin ................................................... 16, 307
Condon Emer M. ........................................... 27, 783
Corron Ned J. ............................................... 34, 1047
Costamagna Eugenio ...................................... 32, 955
Curran Paul .................................................... 22, 27, 30, 547, 783, 895

D
Daafouz Jamal .............................................. 28, 815
Dahlhaus Rainer ............................................ 14, 239
Davis Peter .................................................... 15, 29, 32, 287, 843, 959
De Feo Oscar .................................................. 14, 17, 34, 247, 355, 1039
De Marchi Luca .............................................. 29, 855
De Michele Luca A. ........................................ 24, 635
Di Bernardo Mario .......................................... 22, 25, 555, 659
Di Sciascio Eugenio ........................................ 28, 827
Dmitriev Alexander ........................................ 20, 28, 479, 795
Doi Shinji ....................................................... 10, 23, 71, 591
Duca Anton ..................................................... 34, 1059
Duignan Róisín ............................................... 30, 895
<table>
<thead>
<tr>
<th>Name</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ebina Hiroaki</td>
<td>10, 75</td>
</tr>
<tr>
<td>Efremova Elena</td>
<td>28, 795</td>
</tr>
<tr>
<td>Ehara Takeshi</td>
<td>22, 559</td>
</tr>
<tr>
<td>Endo Tetsuro</td>
<td>13, 179, 183</td>
</tr>
<tr>
<td>Endo Yasunori</td>
<td>28, 30, 787, 887</td>
</tr>
<tr>
<td>Ergun Salih</td>
<td>18, 391</td>
</tr>
<tr>
<td>Hattori Tetsuo</td>
<td>24, 623</td>
</tr>
<tr>
<td>Haurogne Jacky</td>
<td>15, 275</td>
</tr>
<tr>
<td>Hayakawa Yoshihiro</td>
<td>21, 29, 503, 867</td>
</tr>
<tr>
<td>Hayashi Hatsuo</td>
<td>23, 595</td>
</tr>
<tr>
<td>Hayashi Naoki</td>
<td>32, 979</td>
</tr>
<tr>
<td>Hayes Scott T.</td>
<td>34, 1047</td>
</tr>
<tr>
<td>He Jianhua</td>
<td>20, 487</td>
</tr>
<tr>
<td>He Jufang</td>
<td>24, 619</td>
</tr>
<tr>
<td>Heid Christian</td>
<td>23, 607</td>
</tr>
<tr>
<td>Higuchi Kenichi</td>
<td>20, 491</td>
</tr>
<tr>
<td>Hikihara Takashi</td>
<td>10, 31, 75, 927</td>
</tr>
<tr>
<td>Hillier Dàniel</td>
<td>32, 983</td>
</tr>
<tr>
<td>Hiramatsu Akira</td>
<td>34, 1055</td>
</tr>
<tr>
<td>Hirata Yoshito</td>
<td>17, 351, 363</td>
</tr>
<tr>
<td>Hirose Tetsuya</td>
<td>32, 975</td>
</tr>
<tr>
<td>Hisakado Takashi</td>
<td>33, 995</td>
</tr>
<tr>
<td>Hishida Toshihiro</td>
<td>13, 179</td>
</tr>
<tr>
<td>Hoang Thang M.</td>
<td>12, 155</td>
</tr>
<tr>
<td>Hoang Vu</td>
<td>19</td>
</tr>
<tr>
<td>Horai Shunsuke</td>
<td>27, 767</td>
</tr>
<tr>
<td>Horikawa Yo</td>
<td>26, 719</td>
</tr>
<tr>
<td>Horio Yoshihiko</td>
<td>21, 35, 499, 1067, 1079</td>
</tr>
<tr>
<td>Horiuchi Kazuo</td>
<td>11, 131</td>
</tr>
<tr>
<td>Hosokawa Yasuteru</td>
<td>22, 543</td>
</tr>
<tr>
<td>Houssem Jerbi</td>
<td>35, 1071</td>
</tr>
<tr>
<td>Hu SQ</td>
<td>31, 911</td>
</tr>
<tr>
<td>Igarashi Jun</td>
<td>23, 595</td>
</tr>
<tr>
<td>Igunii Youji</td>
<td>26, 703</td>
</tr>
<tr>
<td>Iijima Hiroshi</td>
<td>29, 851</td>
</tr>
<tr>
<td>Ikeda Kenji</td>
<td>30, 903</td>
</tr>
<tr>
<td>Ikeda Yoshikazu</td>
<td>11, 17, 135, 343</td>
</tr>
<tr>
<td>Imura Jun-ichi</td>
<td>15, 267</td>
</tr>
<tr>
<td>Inaba Naohiko</td>
<td>34, 1031</td>
</tr>
<tr>
<td>Inaba Toshio</td>
<td>14</td>
</tr>
<tr>
<td>Inoue Junko</td>
<td>23, 591</td>
</tr>
<tr>
<td>Inoue Kaoru</td>
<td>21, 527</td>
</tr>
<tr>
<td>Inoue Masatomo</td>
<td>26, 731</td>
</tr>
<tr>
<td>Inoue Takahiro</td>
<td>17, 379</td>
</tr>
<tr>
<td>Inuiguchi Masahiro</td>
<td>10, 87, 95</td>
</tr>
<tr>
<td>Ioka Eri</td>
<td>32, 971</td>
</tr>
<tr>
<td>Ishida Yoshihisa</td>
<td>34, 1051</td>
</tr>
<tr>
<td>Ishii Takashi</td>
<td>24, 623</td>
</tr>
<tr>
<td>Ishikawa Yoichi</td>
<td>34, 1051</td>
</tr>
<tr>
<td>Itai Akitoshi</td>
<td>32, 987</td>
</tr>
<tr>
<td>Iu Herbert H.C.</td>
<td>12, 21, 159, 515</td>
</tr>
<tr>
<td>Ivanchenko Mikhail</td>
<td>12, 143</td>
</tr>
<tr>
<td>J</td>
<td></td>
</tr>
<tr>
<td>Jamitzky Ferdinand</td>
<td>31, 915</td>
</tr>
<tr>
<td>Jansson Christian</td>
<td>16, 327</td>
</tr>
<tr>
<td>Jessa Mieczyslaw</td>
<td>25, 691</td>
</tr>
<tr>
<td>Jin’no Kenya</td>
<td>24, 611</td>
</tr>
<tr>
<td>Jitsumatsu Yutaka</td>
<td>28, 803</td>
</tr>
<tr>
<td>Johansson Christopher</td>
<td>23, 599</td>
</tr>
<tr>
<td>K</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Kaleshi Dritan</td>
<td>20, 487</td>
</tr>
<tr>
<td>Kami Takeshi</td>
<td>22, 28, 539, 791</td>
</tr>
<tr>
<td>Kanakov Oleg</td>
<td>11, 127</td>
</tr>
<tr>
<td>Kanaya Ichiro</td>
<td>35, 1075</td>
</tr>
<tr>
<td>Kanazawa Takafumi</td>
<td>12, 167</td>
</tr>
<tr>
<td>Kanemitsu Hideo</td>
<td>30, 875</td>
</tr>
<tr>
<td>Kanno Takahiro</td>
<td>12, 151</td>
</tr>
<tr>
<td>Kanazawa Yuchi</td>
<td>28, 31, 787, 947</td>
</tr>
<tr>
<td>Kao Minghui</td>
<td>15, 291</td>
</tr>
<tr>
<td>Kimihiro Toshiki</td>
<td>21, 527</td>
</tr>
<tr>
<td>Kawakami Hiroshi</td>
<td>27, 747</td>
</tr>
<tr>
<td>Kawakami Shota</td>
<td>23, 571</td>
</tr>
<tr>
<td>Kawamura Arata</td>
<td>26, 703</td>
</tr>
<tr>
<td>Kawamura Satoshi</td>
<td>11, 99</td>
</tr>
<tr>
<td>Kawata Junji</td>
<td>33, 999</td>
</tr>
<tr>
<td>Kennedy Michael P</td>
<td>24, 33, 639, 1003</td>
</tr>
<tr>
<td>Kern Albert</td>
<td>33, 1015, 1019, 1027</td>
</tr>
<tr>
<td>Kim Myoungyoun</td>
<td>19, 427</td>
</tr>
<tr>
<td>Kitahara Hirokazu</td>
<td>27, 767</td>
</tr>
<tr>
<td>Kitajima Hiroyuki</td>
<td>24, 27, 32, 623, 751, 971</td>
</tr>
<tr>
<td>Kletsov Andrey</td>
<td>20, 479</td>
</tr>
<tr>
<td>Kocarev Ljupco</td>
<td>25, 651, 663, 667, 671</td>
</tr>
<tr>
<td>Koida Hiroshi</td>
<td>28, 803</td>
</tr>
<tr>
<td>Kolumban Geza</td>
<td>20, 475</td>
</tr>
<tr>
<td>Komatsu Kazuo</td>
<td>30, 907</td>
</tr>
<tr>
<td>Konishi Keiji</td>
<td>22, 535</td>
</tr>
<tr>
<td>Konishi Takefumi</td>
<td>28, 823</td>
</tr>
<tr>
<td>Konno Hideaki</td>
<td>30, 875</td>
</tr>
<tr>
<td>Kousaka Takuiji</td>
<td>10, 63</td>
</tr>
<tr>
<td>Krebesz Tamas</td>
<td>20, 475</td>
</tr>
<tr>
<td>Krupar Jörg</td>
<td>10, 59</td>
</tr>
<tr>
<td>Krupar Jörg</td>
<td>15, 271</td>
</tr>
<tr>
<td>Kryukov Alexey</td>
<td>9, 55</td>
</tr>
<tr>
<td>Kumagai Sadatoshi</td>
<td>10, 18, 23, 71, 403, 591</td>
</tr>
<tr>
<td>Kumagai Yoshiaki</td>
<td>16, 331</td>
</tr>
<tr>
<td>Kurokawa Hiroyuki</td>
<td>17, 347</td>
</tr>
<tr>
<td>Kurokawa Kaname</td>
<td>13, 211</td>
</tr>
<tr>
<td>Kurosawa Gen</td>
<td>27, 759</td>
</tr>
<tr>
<td>Kurths Jürgen</td>
<td>9, 12, 55, 143</td>
</tr>
<tr>
<td>Kurths Jürgen</td>
<td>14, 235, 239, 243</td>
</tr>
<tr>
<td>Kuzmin Lev</td>
<td>20, 28, 479, 795</td>
</tr>
<tr>
<td>Kwak Kyungsup</td>
<td>23, 575</td>
</tr>
<tr>
<td>Kwok H. S.</td>
<td>25, 687</td>
</tr>
<tr>
<td>Lorenzani Emilio</td>
<td>21, 511</td>
</tr>
<tr>
<td>Lu Jianming</td>
<td>21, 523</td>
</tr>
<tr>
<td>Lunze Jan</td>
<td>10, 59</td>
</tr>
<tr>
<td>Luo Dong X</td>
<td>17, 367</td>
</tr>
<tr>
<td>Luo Xiaodong</td>
<td>17, 359</td>
</tr>
<tr>
<td>Luzzatto Stefano</td>
<td>20, 471</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Ma Kun</td>
<td>21, 507</td>
</tr>
<tr>
<td>Ma Yue</td>
<td>10, 63</td>
</tr>
<tr>
<td>Maeda Michiharu</td>
<td>29, 835</td>
</tr>
<tr>
<td>Maity Somnath</td>
<td>12, 163</td>
</tr>
<tr>
<td>Makino Mitsunori</td>
<td>29, 839</td>
</tr>
<tr>
<td>Marcos Sylvie</td>
<td>12, 147</td>
</tr>
<tr>
<td>Martignoli Stefan</td>
<td>23, 33, 603, 1015</td>
</tr>
<tr>
<td>Martucci Massimiliano</td>
<td>13, 195</td>
</tr>
<tr>
<td>Masuda Seiya</td>
<td>28, 823</td>
</tr>
<tr>
<td>Masugata Osamu</td>
<td>32, 959</td>
</tr>
<tr>
<td>Matarazzo Benedetto</td>
<td>10, 83</td>
</tr>
<tr>
<td>Mathis Wolfgang</td>
<td>10, 67</td>
</tr>
<tr>
<td>Matsueda Hideaki</td>
<td>31, 939</td>
</tr>
<tr>
<td>Matsumoto Akio</td>
<td>14, 231</td>
</tr>
<tr>
<td>Matsumoto Tadashi</td>
<td>13, 18, 191, 407</td>
</tr>
<tr>
<td>Matsuoka Yusuke</td>
<td>34, 1035</td>
</tr>
<tr>
<td>Matsushita Haruna</td>
<td>13, 199, 207</td>
</tr>
<tr>
<td>Mazzini Gianluca</td>
<td>32, 34, 963, 967, 1063</td>
</tr>
<tr>
<td>McAllister Ryan</td>
<td>15, 283</td>
</tr>
<tr>
<td>McGeehan Joe</td>
<td>20, 487</td>
</tr>
<tr>
<td>Meherzi Souyma</td>
<td>12, 147</td>
</tr>
<tr>
<td>Menhaj Mohammad B</td>
<td>12, 139</td>
</tr>
<tr>
<td>Mera Maria E.</td>
<td>27, 771</td>
</tr>
<tr>
<td>Merrkworth Christian</td>
<td>31, 943</td>
</tr>
<tr>
<td>Meucci Riccardo</td>
<td>19, 447</td>
</tr>
<tr>
<td>Mihara Takaya</td>
<td>19, 443</td>
</tr>
<tr>
<td>Millerioux Gilles</td>
<td>28, 815</td>
</tr>
<tr>
<td>Minagawa Junichi</td>
<td>14, 223</td>
</tr>
<tr>
<td>Minagawa T</td>
<td>31, 935</td>
</tr>
<tr>
<td>Misawa Tetsuya</td>
<td>14</td>
</tr>
<tr>
<td>Misuma Masato</td>
<td>29, 839</td>
</tr>
<tr>
<td>Mitsushi Sho</td>
<td>17, 379</td>
</tr>
<tr>
<td>Miyajima Hiromi</td>
<td>29, 835</td>
</tr>
<tr>
<td>Miyakita Kazuyuki</td>
<td>32, 951</td>
</tr>
<tr>
<td>Miyakoshi Masaaki</td>
<td>30, 875</td>
</tr>
<tr>
<td>Miyamoto Sadaaki</td>
<td>28, 30, 787, 887</td>
</tr>
<tr>
<td>Miyamoto Toshiyuki</td>
<td>18, 403</td>
</tr>
<tr>
<td>Miyano Takaya</td>
<td>12, 151</td>
</tr>
<tr>
<td>Mizumura Keisuke</td>
<td>19, 451</td>
</tr>
<tr>
<td>Mizutani Shin</td>
<td>29, 843</td>
</tr>
<tr>
<td>Mocenni Chiara</td>
<td>11, 123</td>
</tr>
<tr>
<td>Molera Lourdes</td>
<td>17, 371</td>
</tr>
<tr>
<td>Mondragon Raul J.</td>
<td>25, 675</td>
</tr>
<tr>
<td>Montefrancesco Giuseppe</td>
<td>11, 123</td>
</tr>
<tr>
<td>Morel Dominique</td>
<td>15, 275</td>
</tr>
<tr>
<td>Mori Kenji</td>
<td>28, 791</td>
</tr>
<tr>
<td>Morie Takashi</td>
<td>21, 495</td>
</tr>
<tr>
<td>Moro Seichiro</td>
<td>13, 18, 191, 407</td>
</tr>
<tr>
<td>Morán Manuel</td>
<td>27, 771</td>
</tr>
<tr>
<td>Muneyasu Mitsuji</td>
<td>26, 711</td>
</tr>
<tr>
<td>Munro Alistair</td>
<td>20, 487</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Muramatsu Jun</td>
<td>15, 29, 287, 843</td>
</tr>
<tr>
<td>Murashima Sadayuki</td>
<td>34, 1055</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naceur Benhadj B</td>
</tr>
<tr>
<td>Nagashima Tomomasa</td>
</tr>
<tr>
<td>Nakaaji Yuki</td>
</tr>
<tr>
<td>Nakagawa Masahiro</td>
</tr>
<tr>
<td>Nakajima Koji</td>
</tr>
<tr>
<td>Nakamura Tomomichi</td>
</tr>
<tr>
<td>Nakamura Yoshimasa</td>
</tr>
<tr>
<td>Nakamura Youhei</td>
</tr>
<tr>
<td>Nakanishi Ryo</td>
</tr>
<tr>
<td>Nakanishi Takahiro</td>
</tr>
<tr>
<td>Nakano Keisuke</td>
</tr>
<tr>
<td>Nakao Hiroya</td>
</tr>
<tr>
<td>Nakao Mitsuhiro T</td>
</tr>
<tr>
<td>Nakashizuka Makoto</td>
</tr>
<tr>
<td>Nakayama Yusuke</td>
</tr>
<tr>
<td>Nassar Elias</td>
</tr>
<tr>
<td>Neto Francisco de Assis Carvalho da Silva</td>
</tr>
<tr>
<td>Nguyen Hung D.</td>
</tr>
<tr>
<td>Nikravesh Sayyed K.</td>
</tr>
<tr>
<td>Nin Jordi</td>
</tr>
<tr>
<td>Ninomiya Kensuke</td>
</tr>
<tr>
<td>Nishi Tetsuo</td>
</tr>
<tr>
<td>Nishida Takaaki</td>
</tr>
<tr>
<td>Nishio Yoshifumi</td>
</tr>
<tr>
<td>Nomura Kentaro</td>
</tr>
<tr>
<td>Nonaka Yasuo</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oancea Servilia</td>
</tr>
<tr>
<td>Obita Yoshinori</td>
</tr>
<tr>
<td>Oda Masayoshi</td>
</tr>
<tr>
<td>Odorizzi Andrea</td>
</tr>
<tr>
<td>Ogata Kenji</td>
</tr>
<tr>
<td>Ogata Masato</td>
</tr>
<tr>
<td>Ogawa Naoki</td>
</tr>
<tr>
<td>Ogita Takeshi</td>
</tr>
<tr>
<td>Ogorzalek Maciej</td>
</tr>
<tr>
<td>Ohashi Yuya</td>
</tr>
<tr>
<td>Ohira Osamu</td>
</tr>
<tr>
<td>Ohnishi Yuiiro</td>
</tr>
<tr>
<td>Ohno Atsuko</td>
</tr>
<tr>
<td>Ohsumi Akira</td>
</tr>
<tr>
<td>Ohta Daisuke</td>
</tr>
<tr>
<td>Ohta Yuzo</td>
</tr>
<tr>
<td>Ohtsuka K.</td>
</tr>
<tr>
<td>Oikawa M.</td>
</tr>
<tr>
<td>Okamura Ryunosuke</td>
</tr>
<tr>
<td>Onozato Yoshikuni</td>
</tr>
<tr>
<td>Osana Yuko</td>
</tr>
<tr>
<td>Osipov Grigory</td>
</tr>
<tr>
<td>Osogami Masahiro</td>
</tr>
<tr>
<td>Osseily Hassan A.</td>
</tr>
<tr>
<td>Otake Tsuyoshi</td>
</tr>
<tr>
<td>Ott Thomas</td>
</tr>
<tr>
<td>Ozaki Katsuhisa</td>
</tr>
<tr>
<td>Ozaki Mitsutoshi</td>
</tr>
<tr>
<td>Ozoguz Serdar</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palumbo Gaetano</td>
</tr>
<tr>
<td>Panajotov Krassimir</td>
</tr>
<tr>
<td>Panas Anrey</td>
</tr>
<tr>
<td>Panayotov Krassimir</td>
</tr>
<tr>
<td>Paoli Andrea</td>
</tr>
<tr>
<td>Papamarkou Theodore</td>
</tr>
<tr>
<td>Pareschi Fabio</td>
</tr>
<tr>
<td>Parlitz Ulrich</td>
</tr>
<tr>
<td>Pennisi Melita</td>
</tr>
<tr>
<td>Pennisi Salvatore</td>
</tr>
<tr>
<td>Petheil Shawn D.</td>
</tr>
<tr>
<td>Phothisonothai Montri</td>
</tr>
<tr>
<td>Plum Michael</td>
</tr>
<tr>
<td>Polovinkin Andre</td>
</tr>
<tr>
<td>Prachumrak Korakot</td>
</tr>
<tr>
<td>Probst Florian A.</td>
</tr>
<tr>
<td>Putkaradze Vakhtang</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qiu Shui-Sheng</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramberto Marco</td>
</tr>
<tr>
<td>Rakoto-Ravalontsalama Naly</td>
</tr>
<tr>
<td>Raman A.</td>
</tr>
<tr>
<td>Rezgui Nizar</td>
</tr>
<tr>
<td>Righero Marco</td>
</tr>
<tr>
<td>Rinaldi Sergio</td>
</tr>
<tr>
<td>Rocchi Santina</td>
</tr>
<tr>
<td>Romano Maria C.</td>
</tr>
<tr>
<td>Rovatti Riccardo</td>
</tr>
<tr>
<td>Roy Rajarshi</td>
</tr>
<tr>
<td>Rugin Raffaele</td>
</tr>
<tr>
<td>Rump Siegfried M.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saboori Iman</td>
</tr>
<tr>
<td>Sagara Nobusumi</td>
</tr>
<tr>
<td>Saito Toshimichi</td>
</tr>
<tr>
<td>Sakamoto Masaki</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>------------------------</td>
</tr>
<tr>
<td>Sakamoto Masashi</td>
</tr>
<tr>
<td>Sakamoto Takeshi</td>
</tr>
<tr>
<td>Salvadori Francesco</td>
</tr>
<tr>
<td>Samadian Sohrab</td>
</tr>
<tr>
<td>Sandra Pavel</td>
</tr>
<tr>
<td>Sano Satoshi</td>
</tr>
<tr>
<td>Sato Akari</td>
</tr>
<tr>
<td>Sato Kosuke</td>
</tr>
<tr>
<td>Sato Masatoshi</td>
</tr>
<tr>
<td>Sato Masayuki</td>
</tr>
<tr>
<td>Sato Shigeo</td>
</tr>
<tr>
<td>Savazzi Pietro</td>
</tr>
<tr>
<td>Sawahashi Mamoru</td>
</tr>
<tr>
<td>Schelter Björn</td>
</tr>
<tr>
<td>Schiller Axel</td>
</tr>
<tr>
<td>Schilder Frank</td>
</tr>
<tr>
<td>Schmackers Judith</td>
</tr>
<tr>
<td>Schwarz Wolfgang</td>
</tr>
<tr>
<td>Sciamanna Marc</td>
</tr>
<tr>
<td>Sebatu Carla</td>
</tr>
<tr>
<td>Secchi Cristian</td>
</tr>
<tr>
<td>Sekikawa Munehisa</td>
</tr>
<tr>
<td>Sekiya Hiro</td>
</tr>
<tr>
<td>Sengoku Masakazu</td>
</tr>
<tr>
<td>Setti Gianluca</td>
</tr>
<tr>
<td>Shigeti Noritaka</td>
</tr>
<tr>
<td>Shima Daisuke</td>
</tr>
<tr>
<td>Shimada Tatsuro</td>
</tr>
<tr>
<td>Shimbo Masaru</td>
</tr>
<tr>
<td>Shimizu Kuniyasu</td>
</tr>
<tr>
<td>Shimizu Yoshiaki</td>
</tr>
<tr>
<td>Shinoda Shoji</td>
</tr>
<tr>
<td>Shirahama Hiroyuki</td>
</tr>
<tr>
<td>Shirak Yoshinno</td>
</tr>
<tr>
<td>Sievers Albert J.</td>
</tr>
<tr>
<td>Sigmund Karl</td>
</tr>
<tr>
<td>Sinyakin Vladimir</td>
</tr>
<tr>
<td>Small Michael</td>
</tr>
<tr>
<td>Someya Hiroyuki</td>
</tr>
<tr>
<td>Sorrentino Francesco</td>
</tr>
<tr>
<td>Speciale Nicolò</td>
</tr>
<tr>
<td>Stark Robert</td>
</tr>
<tr>
<td>Stellardo Daniele</td>
</tr>
<tr>
<td>Stoop Norbert</td>
</tr>
<tr>
<td>Stoop Ruedi</td>
</tr>
<tr>
<td>Storace Marco</td>
</tr>
<tr>
<td>Suenaga Shinya</td>
</tr>
<tr>
<td>Susuki Yoshihiko</td>
</tr>
<tr>
<td>Suykens Johan A. K.</td>
</tr>
<tr>
<td>Suzuki Hideyuki</td>
</tr>
<tr>
<td>Suzuki Mami</td>
</tr>
<tr>
<td>Suzuki Satoshi</td>
</tr>
<tr>
<td>Taguchi Takaaki</td>
</tr>
<tr>
<td>Taira Hiroto</td>
</tr>
<tr>
<td>Takahashi Naoyuki</td>
</tr>
<tr>
<td>Takahashi Nobuaki</td>
</tr>
<tr>
<td>Takahashi Norikazu</td>
</tr>
<tr>
<td>Takai Shigemasa</td>
</tr>
<tr>
<td>Takata Hitoshi</td>
</tr>
<tr>
<td>Takumi Ichiro</td>
</tr>
<tr>
<td>Tamura Hiroshi</td>
</tr>
<tr>
<td>Tan Kangrong</td>
</tr>
<tr>
<td>Tanabe Kunio</td>
</tr>
<tr>
<td>Tanaka Hideki</td>
</tr>
<tr>
<td>Tanaka Hidenori</td>
</tr>
<tr>
<td>Tanaka Hisa-Aki</td>
</tr>
<tr>
<td>Tanaka Mamoru</td>
</tr>
<tr>
<td>Tang Kwok W.</td>
</tr>
<tr>
<td>Tang Wallace</td>
</tr>
<tr>
<td>Tanino Tetsuzo</td>
</tr>
<tr>
<td>Tanji Yuichi</td>
</tr>
<tr>
<td>Tanoue Takashi</td>
</tr>
<tr>
<td>Taoka Hidekazu</td>
</tr>
<tr>
<td>Tateno Katsumi</td>
</tr>
<tr>
<td>Tatsumi Keiji</td>
</tr>
<tr>
<td>Teglio Andrea</td>
</tr>
<tr>
<td>Thiel Marco</td>
</tr>
<tr>
<td>Thienpont Hugo</td>
</tr>
<tr>
<td>Tiezzi Silvia</td>
</tr>
<tr>
<td>Timmer Jens</td>
</tr>
<tr>
<td>Titchener Mark R.</td>
</tr>
<tr>
<td>Tokinaga Shozo</td>
</tr>
<tr>
<td>Tomashevskiy Alexey I.</td>
</tr>
<tr>
<td>Tomescu Florian</td>
</tr>
<tr>
<td>Tomita Masato</td>
</tr>
<tr>
<td>Torikai Hiroyuki</td>
</tr>
<tr>
<td>Torra Vicenç</td>
</tr>
<tr>
<td>Tovar Gessyca M.</td>
</tr>
<tr>
<td>Tse Chi K.</td>
</tr>
<tr>
<td>Tsubone Tadashi</td>
</tr>
<tr>
<td>Tsuchie Yoshiyuki</td>
</tr>
<tr>
<td>Tsuji H.</td>
</tr>
<tr>
<td>Tsuji Shigeki</td>
</tr>
<tr>
<td>Tsumoto Kunchiha</td>
</tr>
<tr>
<td>Tsuneda Akio</td>
</tr>
<tr>
<td>Tsurumi Masayo</td>
</tr>
</tbody>
</table>

**T**

<table>
<thead>
<tr>
<th>Name</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tada Yoshifumi</td>
<td>11, 107</td>
</tr>
<tr>
<td>Tagami Takanori</td>
<td>30, 903</td>
</tr>
</tbody>
</table>