Akiyama’s Math Spectacle Show has intrigued and inspired people in numerous countries around the world. Not only has he been spreading the intellectual side of mathematics and its usefulness to the general public, but his lectures also have a spirit of entertainment filled with humor and wit. The specific topics which will be delivered in the lecture are listed in the followings:

1. Create jigsaw puzzles from a regular tetrahedron.
2. Is this a rabbit or a duck?
3. A bird in a cage.
4. Spider changes to Geisha Girl.
5. It is a donkey or a dog? (Fig.3)
6. A big snake swallows a fox.
7. A pig does gymnastics on a horizontal bar.
8. Who is in the U.F.O?
9. Can two giant pandas live in the same house together? (Fig.1)
10. One type of atom fills space.
11. Why is a manhole cover round?
12. Can you drill a triangular hole?
13. Vehicles with strange wheels like bagels flattened along the edges.
14. How to pack cans efficiently?
15. Soap tells us is the minimum network connecting many cities.
16. A compact disk (CD) works well even if it is damaged.
17. Kidney stone will be cured without operation.
18. Volume and surface area of a sphere. (Fig.2)
20. Pythagorean Theorem and its applications.

Fig. 1  
Fig.2
References

Akiyama’s Profile

Jin Akiyama is a mathematician at heart. Currently, he is a fellow of the European Academy of Science, the founding editor of the journal Graphs and Combinatorics, and the director of the Research Institute of Math Education at Tokyo University of Science. He is particularly interested in graph theory and discrete and solid geometry, and has published many papers in these fields. Aside from his research, he is best known for popularizing mathematics, first in Japan and then in other parts of the globe: his lecture series was broadcast on NHK television and radio from 1991 to 2013. He was a founding member of the Organizing Committee of the UNESCO-sponsored traveling exhibition “Experiencing Mathematics”, which debuted in Denmark in 2004. In 2013, he built a hands-on mathematics museum called “Akiyama’s Math Experience Plaza” in Tokyo. He has authored and co-authored more than 200 books, including *Factors and Factorizations of Graphs* (jointly with M. Kano, Lect. Notes Math., Springer, 2011) and, *A Day’s Adventure in Math Wonderland* (jointly with M. Ruiz, World Scientific, 2008), which has been translated into nine languages. He was recently awarded the book prize by Japan Math Society.
Detecting Bifurcations from Time Series Data

Kazuto OGOSHI, Hiroyuki KITAJIMA and Toru YAZAWA

†Faculty of Engineering, Kagawa University
2217-20 Hayashi, Takamatsu, Kagawa 761-0396, Japan
Email: s16g454@stu.kagawa-u.ac.jp, {kitaji,tyazawa}@eng.kagawa-u.ac.jp

Abstract—In this paper, we try to detect the early-warning signal from time series data. Simple mathematical models with noise and a real system are used to obtain the data. We calculate autocorrelation functions at the parameter values close to and away from a bifurcation. We determine that an averaged peak interval of the autocorrelation function is a good indicator for predicting bifurcations.

1. Introduction

Detecting bifurcation points from time series is the important and challenging issue because complex dynamical systems, ranging from ecosystems to financial markets and the climate, can have critical points at which a sudden shift to a contrasting dynamical regime may occur [1, 2, 3]. Scheffer et al. reviewed many methods of detecting early-warning signals applied to simple mathematical models to real systems [1]. Chen et al. detected a “pre-disease” state using network biomarkers [4]. Peng et al. proposed detrended fluctuation analysis (DFA) for determining the statistical self-affinity of a signal [5] and applied it to heartbeat time series data [6]. Yazawa proposed a modified DFA method and obtained a scaling index of judging a healthy condition from heartbeat [7].

In this study, we use the autocorrelation function (ACF) to capture early-warning signals just before bifurcations in simple mathematical models. Then, we apply it to heartbeat data from an experiment. We determine that the averaged peak interval of the autocorrelation function is one of good indicators for detecting bifurcations.

2. Model equations

We use the BVP model [8] and Luo-Rudy model [9] in this study. BVP equations are described by

\[
\begin{align*}
\frac{dx}{dt} &= \omega y - \sigma x \\
\frac{dy}{dt} &= -\omega x + \epsilon(1 - \beta y^2)y.
\end{align*}
\]  

These equations describe the behavior of an electric circuit containing an inductor, a capacitor, and linear and nonlinear resistors [8]. These equations are derived from the simplification of the Hodgkin-Huxley equations which describe the electric property of the squid axon. In this paper, we set \( \omega = 1.0 \) and \( \beta = 1.0 \), and change the values of \( \epsilon \) and \( \sigma \).

The membrane potential \( V \) of the LR model with the synaptic external input is described by

\[
C \frac{dV}{dt} = -(I_{Na} + I_{Ca} + I_K + I_{K1} + I_{Kp} + I_b + I_{syn}).
\]

This is a mathematical model of the mammalian (guinea pig) ventricular cell. In this paper, we change the values of \([K]_o \) (free concentration of the potassium ions in the extracellular compartment) because we had determined that increasing \([K]_o \) triggers a period-doubling bifurcation generating alternans [10].

3. Results

3.1. BVP model

A bifurcation diagram is shown in Fig. 1. In this figure, thin solid and dotted curves indicate Hopf and pitchfork bifurcations of equilibrium points, respectively. A periodic solution is generated by crossing the Hopf bifurcation from bottom to up. The periodic solution disappears by the tangent bifurcation denoted by a thick solid curve in Fig. 1. We observe the stable periodic solution in a gray parameter region. Here, we try to quantify the stability of the periodic solution from time series of the state variable \( x \). We set \( \epsilon = 2.0 \) and change the value of \( \sigma \) in Eq. (1) to control the distance to the tangent bifurcation.

Figures 2(a) and 2(b) show waveforms of \( x \) at the points away from and close to the bifurcation point. Next, we
add the Gaussian white noise $\xi(t)$ to $x$ in Eq. (1), where $\langle \xi(t) \xi(t') \rangle = \delta(t-t')$. Noise-added waveforms of $x$ are shown in Figs. 3(a) and 3(b). We calculate ACFs for Figs. 3(a) and 3(b). The ACF of data $X_i$ is given by
\[
R(\tau) = \frac{E[(X_i - \mu)(X_{i+\tau} - \mu)]}{\eta^2}
\]
where $E[\cdot]$ is the expected value, $\mu$ and $\eta$ are the mean and variance of data $X_i$, respectively. The results are shown in Figs. 4(a) and 4(b). Next, we calculate the mean of peak intervals of low-passed noise-added waveforms as a function of the parameter $\sigma$ to estimate the stability of the periodic solution. The result is shown by a solid curve in Fig. 5.

![Figure 2: Waveforms of $x$](image)

![Figure 3: Noise-added waveforms of $x$](image)

![Figure 4: Autocorrelation function of noise-added waveforms](image)

![Figure 5: Period of periodic solution (dotted curve) and mean of peak intervals of noise-added waveform (solid curve) as a function of $\sigma$](image)

3.2. Luo-Rudy model

Next, we apply our indicator to the Luo-Rudy model. We try to quantify the stability of the periodic solution from time series of the membrane potential $V$. We change the
values of [K]₀: a normal value in [9] is 5.4 and a period-
doubling bifurcation occurs around 6.583 [10]. Figures
6(a) and 6(b) show waveforms of V at these parameter val-
ues. Next, we add the Gaussian white noise to V in Eq. (2).
We add the noise during phase 2 (plateaus of V). Resultant
waveforms are shown in Figs. 7(a) and 7(b). We calculate
ACFs for these waveforms, which are shown in Figs.
8(a) and 8(b). A two-periodic-like waveform of the ACF
appears only in Fig. 8(b), thus the ACF has a possibility of
discrimination between Figs. 6(a) and 6(b).

![Waveforms of V in Luo-Rudy model](image1)

Figure 6: Waveforms of V in Luo-Rudy model

![Noise-added waveforms for Luo-Rudy model](image2)

Figure 7: Noise-added waveforms for Luo-Rudy model

3.3. Experimental data

Next, we apply our indicator to a real system. Yazawa
measured the heartbeat of bumblebees for a long time. Fig-
ures 9(a) and 9(b) show typical waveforms of a normal state
and just before generating abnormal rhythm. We apply the
previous process except for adding noise to these data be-
cause these measured data already contain noise. Figures
10(a) and 10(b) show ACFs of Figs. 9(a) and 9(b). The
results of the averaged peak intervals are 143.999[ms] and
219.857[ms] for a normal state and just before an abnor-
mal state, respectively. The indicator is increased as the
state approaches a abnormal state. It is an opposite result
to the BVP model. The detailed analysis is one of our fu-
ture problems.

4. Conclusion

We investigated indicators of detecting bifurcation
points from time series in mathematical models and exper-
imental data. First, we determined that the averaged peak
interval of the autocorrelation function is one of good in-
dicators for detecting a tangent bifurcation using the BVP
model. Second, we applied our indicator to more realistic
mathematical model: a model of the mammalian (guinea
pig) ventricular cell. The autocorrelation function between
a normal state and just before a period-doubling bifurcation
have different shapes, however the averaged peak interval
of the autocorrelation function is almost the same because
the period of the external input is the same. We are now
trying to quantify these differences. Last, we applied it to
experimental data on heartbeat of bumblebees. However,
the result is opposite to the BVP’s case. We need more
results for various computer simulations and experimental
data to check the validity of our method. Now we are tackling such issues to obtain better results.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K00405.

References

Bifurcation Analysis for Early Afterdepolarization in Shannon Model

Ai Takahashi, Hiroyuki Kitajima and Toru Yazawa

Résumé—La membrane potentielle du myocyte ventriculaire est régulée par des courants ioniques entrants et sortants. Lorsque ces activités électriques sont perturbées, un arrêt cardiaque se produit. Dans un cas particulier, la prolongation QT est considérée comme un signe de cessation cardiaque soudaine. Dans cette étude, nous examinons la relation entre les courants ioniques et l'apparition d'arrêts cardiaques. Nous obtenons que la prolongation QT est liée à la non-linearité des courants sodium et calcium et à l'occurrence d'arrêts cardiaques. Ce phénomène est décrit par la méthode de bifurcation. En appliquant la méthode de bifurcation à ce modèle, nous obtenons que l'augmentation du courant sodium-calcium influence la prolongation QT et les EADs.

1. Introduction

Les activités électriques sont en cours dans la membrane cellulaire des myocytes ventriculaires. Cela servira à régler les contractions cardiaques régulées. L'activité électrique est causée par l'action potentielle. Une diagramme d'une action potentielle cardiaque typique est montré dans la figure 1. L'action potentielle se réfère à la réaction d'excitation de la cellule avec la dépolarisation. Après la membrane potentielle est dépolarisée (approchant 0 [mV]), l'action potentielle se produit lorsqu'elle atteint le seuil de la membrane potentielle. Le potentiel cardiaque est classé en 5 phases : phase 0 (dépolarisation), phase 1 (spike), phase 2 (plateau), phase 3 (repolarisation) et phase 4 (potentiel de repos). Quand ces activités électriques sont perturbées, un arrêt cardiaque se produit. Ce phénomène est dû à la membrane potentielle. La membrane potentielle devient faible comme −40 [mV]. Si la membrane potentielle est faible, l'arrêt cardiaque est étonnant. Il devient la cause de l'excitation cardiaque.

Parmi les activités électriques anormales du myocyte cardiaque, nous payons attention à la QT longue (LQTS). LQTS est un phénomène qui produit de polymorphes de la tachycardie à pointes (TdP) et prolonge la QT interval dans l'électrocardiogramme. LQTS est lié à des afterdepolarizations (AEDs). LQTS peut aboutir à un arrêt cardiaque soudain en raison de la fibrisation ventriculaire. Les LQTS sont classifiés en 13 types [1]. Les types les plus communs sont LQT 1, 2 et 3. Il est dit que LQT 1 et 2, et 3 sont causés par l'aberration de calcium et du sodium ion channels, respectivement [2, 3, 4, 5]. Récemment, un nouveau et sélectif inhibiteur (SEAO400) de l'exchanger sodium-calcium (NCX) a été détecté [6] et son effet sur la génération de LQTS a été étudié [7, 8, 9, 10]. De plus, en utilisant la méthode de bifurcation, nous avons obtenu que l'EDA est causé par la bifurcation homocline dans le modèle Luo-Rudy I (LRI) [11]. Cependant, le LRI model ne comprend pas le NCX courant.

Dans ce papier, nous utilisons le modèle Shannon [12] qui décrit les détails dynamiques du calcium intracellulaire. En appliquant la méthode de bifurcation au modèle Shannon, nous obtenons que le calcium action potential cardiaque prolonge la génération QT et les EADs.

2. Model

Le modèle Shannon décrit les détails dynamiques du calcium dans le ventricule myocyte. Dans ce modèle, quatre compartiments sont considérés : le réticulum sarcoplasmique (SR), le cleft intercellulaire pour SR, l'espace sous-sarcoïdémique, et l'espace cytoplasmique. Le calcium dynamiques est décrivant par le L-type Ca canaux, la Ca leak, le Ca pump, le NCX, et la buffering. Le potentiel de membrane est donné par

\[
C \frac{dV}{dt} = -\left( I_{Na} + I_{NB} + I_{NaK} + I_{Kr} + I_{Ks} + I_{bas} + I_{oFr} + I_{K1} + I_{NaCa} + I_{Ca} + I_{Ch} + I_{CaL} + I_{CaP} + I_{CaX} \right)
\]

où \(V\) est le potentiel cellulaire, \(C\) est la capacité de membrane cellulaire, et \(I_{j}\) sont les courants ioniques sauf pour les synaptic.
current $I_{syn}$. All ionic currents in this model are shown in Table 1. $I_{syn}$ is given by

$$I_{syn} = G_{syn}(V - V_{syn})s(t'),$$

where $G_{syn}$ is the maximum synaptic conductance, $V_{syn}$ is the reversal potential, and $s(t')$ is given by

$$s(t') = \frac{\tau_1}{\tau_2 - \tau_1} \left( - \exp\left(-\frac{t'}{\tau_1}\right) + \exp\left(-\frac{t'}{\tau_2}\right) \right),$$

where $\tau_1$ and $\tau_2$ are the raise and decay time of the synapse. We identify these values ($\tau_1 = 5.5$ and $\tau_2 = 90.0$[ms]) from the experimental data [13]. $t'$ is the time that is reset at every $nT$ ($n$ is a natural number, and $T$ is the BCL: basic cycle length). We check the periodicity of the trajectory by using the state variables at every $nT$. The values of the parameters related with the synapse are fixed as $G_{syn} = 4.0$ and $V_{syn} = -29$.

Some ionic currents have the following form

$$I_j = G_j \cdot y \cdot (V - E_j),$$

where $G_j$ is the maximum conductance and $E_j$ is the reversal potentials for ion $j$. The gating variable $y$ is given by

$$\frac{dy}{dt} = \frac{y_{\infty} - y}{\tau_y},$$

where $\tau_y$ and $y_{\infty}$ are time constant and the value of $y$ in the steady state, respectively. There are 14 gating variables in the Shannon model. The kind of state variables is shown in Tab 2. In total, the Shannon model is described by 39-dimensional ordinary differential equations.

<table>
<thead>
<tr>
<th>Table 1: Ionic currents in Shannon model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Abbreviation</strong></td>
</tr>
<tr>
<td>$I_{Na}$</td>
</tr>
<tr>
<td>$I_{NaK}$</td>
</tr>
<tr>
<td>$I_{NaK}$</td>
</tr>
<tr>
<td>$I_{CaL}$</td>
</tr>
<tr>
<td>$I_{CaB}$</td>
</tr>
<tr>
<td>$I_{CaP}$</td>
</tr>
<tr>
<td>$I_{NaCa}$</td>
</tr>
<tr>
<td>$I_{ClCa}$</td>
</tr>
<tr>
<td>$I_{Kr}$</td>
</tr>
<tr>
<td>$I_{Ks}$</td>
</tr>
<tr>
<td>$I_{so}$</td>
</tr>
<tr>
<td>$I_{so}$</td>
</tr>
<tr>
<td>$I_{K1}$</td>
</tr>
<tr>
<td>$I_{Clb}$</td>
</tr>
</tbody>
</table>

### 3. Results

There are 14 ionic currents in the Shannon model. We investigate the influence of changing the value of each current (multiplying $I_j$ by artificial parameter $Z_j$) on generation of QT prolongation and EADs. We study that the interval of external stimulus (BCL) equals 1000 [ms]. As a result, arrhythmia is observed when each of two ionic currents (sodium-calcium exchange current $I_{NaCa}$ and L-type calcium channel current $I_{CaL}$) is changed. We calculate a two-parameter bifurcation diagram and clarify the dominant parameter for generating QT prolongation and EADs.

We show a two-parameter bifurcation diagram in Fig. 2. In this diagram, $Z_{CaL}$ and $Z_{NaCa}$ are artificial parameters of $I_{CaL}$ and $I_{NaCa}$, respectively. The solid curves denoted by $N$ and $I$ indicate Neimark-Sacker and period-doubling bifurcations, respectively. In the region colored by gray, we observe a normal waveform of the membrane potential as shown in Fig. 3. The neuron fires at every $1000 \times n$ [ms], where $n$ is a natural number. Here, 1000 means the period of the external stimulus. This normal state suddenly disappears by crossing the bifurcation curves.

To show responses after crossing the bifurcation, we use one-parameter bifurcation diagrams as shown in Fig. 4. Figures 4(a) and 4(b) are obtained by changing the parameter values along the arrows (a) and (b) in Fig. 2, respectively. The horizontal axis indicates the peaks of the membrane potential after transient time. We observe one point until $Z_{NaCa} \approx 3.85$ and $Z_{CaL} = 1.55$, which means the membrane potential of the cardiac muscle has one peak during one external stimulus. Roughly speaking, we also observe that the points are on three lines: the top ($V \approx 33$), the middle ($V \approx 0$), and the bottom ($V \approx -35$). We show waveforms of $V$ near bifurcation points in Figs. 5(a) and 5(b). The points on the bottom and middle in Fig. 4 indicate small and large peaks shown by the arrow in Figs. 5(a) and 5(b), respectively. Note that the appearance of the bottom points does not correspond to a bifurcation, it is just the transform of the waveform. On the other hand, the emergence of the middle points, which means the emergence of an EAD, corresponds to a bifurcation; the type of a bifurcation (Neimark-Sacker or period-doubling) depends on the parameter value. EADs occur with abnormal depolarization during phase 2 or phase 3, and are caused by an in-
crease in the frequency of abortive action potentials before normal repolarization is completed [15].

Figure 6 is obtained by changing the parameter values along the arrow (c) ($Z_{NaCa} = 2$) in Fig. 2. The horizontal axis is the same as that of Fig. 4. The Neimark-Sacker bifurcation occurs at $Z_{CaL} \approx 2$, and EADs appear as shown in Fig. 7(a). We can see a plurality of projections and prolongation of the action potential duration (APD), which correspond to QT prolongation in the electrocardiogram. In Fig. 7(b) we show a membrane potential waveform in case of $Z_{NaCa} = 2$ and $Z_{CaL} = 3.5$. We can see the irregular wave. This membrane potential waveform is extremely dangerous.
4. Conclusion

In this study, we examined dependence of generating EADs and QT prolongation on parameter values in the Shannon mathematical model. We calculated a two-parameter bifurcation diagram. As a result, two ionic currents (sodium-calcium exchange current $I_{NaCa}$ and L-type calcium channel current $I_{CaL}$) are keys to the generation of EADs and QT prolongation. We determined that a normal state becomes unstable by bifurcations. Before a bifurcation, the action potential duration (APD) becomes a little longer as the $I_{NaCa}$ or $I_{CaL}$ is increased. After a bifurcation, EADs suddenly appear as a stable state and at the same time the APD becomes very long which corresponds to QT prolongation. Extremely dangerous waveforms like ventricular fibrillation were obtained after occurrence of QT prolongation. This agrees with that QT prolongation is a sign of sudden cardiac cessation.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K00405. We thank Prof. T. Yoshinaga of Tokushima University for providing his powerful bifurcation analysis tools.

References


Analysis of a 3-dimensional piecewise-constant chaos generator
without constraint

Shuntaro Hamatani and Tadashi Tsubone

†Department of Electrical, Electronics and Information Engineering, Nagaoka University of Technology, 
1603-1 Kamitomioka, Nagaoka, 940-2137, Japan 
Email: s143171@stn.nagaokaut.ac.jp, tsubone@vos.nagaokaut.ac.jp

Abstract—In this study, we propose an novel autonomous 3-dimensional piecewise-constant dyna-
nical system without constraint, and analyze the nonliner phenomena by using two-dimensional(2-
D) return map. The return map is derived rigorously and is represented by explicit expressions. In ad-
dition, some experimental results are obtained in the extremely simple circuit.

1. Introduction

The piecewise-constant systems[1] exhibit many nonliner phenomena, chaos, bifurcation and so on, in spite of the simple dynamics. The systems can be analysed with comparatice ease, because the piecewise-solutions are linear and the connections of the solutions are described explity on the bound-
ary[2]. Therefore, the piecewise-constant systems are useful for demonstrating nonlinear phenomena. For example, regorous analysis of quasiperiodic bi-
furcation in piecewise-constant systems with exter-
nal forces have been reported[3][4], and analysis of synchronization in a coupled piecewise-constant system have been discussed [4]. In addition, methods to derive rigorous solutions have been de-
veloped for bath autonomous and non-autonomous piecewise-constant systems[4][5]. However, pre-
vious piecewise-constant circuits are discribed by constrained equations, that is, state variables are constrained in partial hyperplanes of phase space depended on some conditions. Therefore, some considerations are insufficient about more natu-
ral systems without constraint. In this study, we present a novel three-dimensional(3-D) autonomous 

2. 3-D chaos-generationg piecewise-constant os-
cillator

Figure 1 shows the circuit diagram of the piecewise-constant circuit. This circuit consists of six voltage-controlled current sources(VCCSs) that has signum characteristic and three capacitors. These VCCSs are realized by operational transcon-
ductance amplifiers. In order to describe the dynam-
ics, we define two functions as follows:

\[
\text{Sgn}(x) = \begin{cases} 
1 & \text{for } x \geq 0 \\
-1 & \text{for } x < 0 
\end{cases}, 
U(x) = \begin{cases} 
1 & \text{for } x \geq 0 \\
0 & \text{for } x < 0. 
\end{cases}
\]  

(1)

Then, the circuit dynamics is described as follows:

\[
\begin{align*}
\dot{x} &= \text{Sgn}(y - 1) \\
\dot{y} &= -\text{Sgn}(x) - b \cdot \text{Sgn}(y) + a \cdot U(z) \cdot \text{Sgn}(y) \\
\dot{z} &= -\text{Sgn}(x). 
\end{align*}
\]  

(2)

where “ · ”represents the derivative of \( \tau \) and the following normalized variables and parameters are used:

\[
\tau = \frac{I_3}{C_3 E}t, \quad x = \frac{C_1}{C_3 E}v_1, \quad y = \frac{C_1}{C_2 E}v_2, \quad z = \frac{1}{E}v_3, \\
a = \frac{I_a}{I_1}, \quad b = \frac{I_b}{I_2} 
\]  

(3)

\( I_1, I_a \)and \( I_b \) are bias currents of operational transcon-
ductance amplifiers that are represented by trapezoid symbols in Fig.1.
The dynamics are represented by twelve local vector fields and the conditions. These are shown in Table 1.

Fig 1: Implemented circuit \((x \propto v_1, y \propto v_2, z \propto v_3)\)

(a) \(x - y\)  (b) \(x - z\)  (c) \(y - z\)

(d) \(v_1 - v_2\)  (e) \(v_1 - v_3\)  (f) \(v_2 - v_3\)

Fig 2: Typical chaotic attractor \((a=1.5, b=0.7)\)
(a)(b)(c) Regorous solutions , (d)(e)(f) Laboratory measurements 2.0 [V/div.]

Figure 2 shows a typical chaotic attractor with \(a = 1.5, b = 0.7\). We measured same attractor in laboratory.

3. 2-D return map

In order to analyze the chaotic behavior, we focus on parameter \(a = 1.5, b = 0.4\) for simplicity and derive 2-D return map. First, we define the domain \(S\):
\[
S = \{x = (x, y, z) \mid y = 1\}. \tag{4}
\]
The trajectory starting from \(x_0\) on \(S\) must return to \(x_1\) on \(S\). Then, we can define a 2-D return map \(F\) from \(S\) to itself.
\[
F : S \rightarrow S, (x_1, z_1) = F_i(x_0, z_0) = (f_i(x_0, z_0), g_i(x_0, z_0)) \quad (i = 0, 1, 2, \ldots, 7). \tag{5}
\]
There are eight kind of trajectories that from \(S\) to itself. The trajectories are derived by the thresholds that are given as follows:
\[
\begin{align*}
T_{h0} &= \frac{b^2 - 1}{2b(b + 1)} \cdot z_0, \\
T_{h1} &= \frac{1}{1 + b}, \\
T_{h2} &= \frac{(b^3 + b^2 - b - 1) \cdot z_0 - 4 \cdot b}{2 \cdot b^3 - 2 \cdot b}. \tag{6}
\end{align*}
\]

Figure 3 shows the local regions of maps \(F_i\). By using solution of (2), 2-D return map is derived rigorously and is represented by explicit expressions.
\[
(x_1, z_1) = F_i(x_0, z_0)
\]
\[ F_0(x_0, z_0) \quad \text{for} \quad x_0 \geq z_0, x_0 < 0, z_0 < 0, \]
\[ F_1(x_0, z_0) \quad \text{for} \quad x_0 < z_0, x_0 < 0, z_0 < 0, \]
\[ F_2(x_0, z_0) \quad \text{for} \quad x_0 \geq Th_2, x_0 \geq Th_1, z_0 < 0, \]
\[ F_3(x_0, z_0) \quad \text{for} \quad x_0 < Th_2, x_0 \geq Th_1, z_0 < 0, \]
\[ F_4(x_0, z_0) \quad \text{for} \quad x_0 < 0, z_0 \geq 0, \]
\[ F_5(x_0, z_0) \quad \text{for} \quad x_0 \geq Th_0, 0 \leq x_0 < Th_1, \]
\[ F_6(x_0, z_0) \quad \text{for} \quad x_0 < Th_0, 0 \leq x_0 < Th_1, \]
\[ F_7(x_0, z_0) \quad \text{for} \quad x_0 \geq 0, z_0 \geq 0, \]

where \((f_i, g_i)\) are shown in Table 2.

\[ DF_0(x, z) = \begin{bmatrix} -\frac{b-1}{b-1} & 0 \\ \frac{b-1}{b-1} & 1 \end{bmatrix}, \]

for \( x_0 \geq z_0, x_0 < 0, z_0 < 0, \)

\[ DF_1(x, y) = \begin{bmatrix} -\frac{b+2a-1}{b-1} & -\frac{b+2a-1}{b-1} \\ \frac{b+2a-1}{b-1} & 1 - \frac{b+2a-1}{b-1} \end{bmatrix}, \]

for \( x_0 < z_0, x_0 < 0, z_0 < 0, \)

\[ DF_2(x, y) = \begin{bmatrix} \frac{b-1}{b-1} & 0 \\ -\frac{b+1}{b-1} & 1 \end{bmatrix}, \]

for \( x_0 \geq Th_2, x_0 \geq Th_1, z_0 < 0. \) (8)

By using these Jacobian matrices, we obtained the eigenvalues of Jacobian matrix of the \(n\) times composition map for large \(n\) theoretically. Since any one of the absolute values of eigenvalues is \(|\lambda| \geq 1\), we can say that the attractor shown in Fig. 4 is unstable.

4. Conclusion

We realized 3-D autonomous piecewise-constant chaos generator without constraint. Using 2-D return map, we confirmed the unstability of the chaotic behavior. In the future, we will clarify the existence region of chaotic attractor.

References


Table 1: Local vector fields and regions for $k$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$\alpha(k)$</th>
<th>$D_k$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$t(-1, 1+b, 1)$</td>
<td>$x</td>
</tr>
<tr>
<td>1</td>
<td>$t(-1, -1+b, -1)$</td>
<td>$x</td>
</tr>
<tr>
<td>2</td>
<td>$t(-1, 1+b, 1)$</td>
<td>$x</td>
</tr>
<tr>
<td>3</td>
<td>$t(-1, -1-b, -1)$</td>
<td>$x</td>
</tr>
<tr>
<td>4</td>
<td>$t(1, 1-b, 1)$</td>
<td>$x</td>
</tr>
<tr>
<td>5</td>
<td>$t(-1, -1-b, -1)$</td>
<td>$x</td>
</tr>
<tr>
<td>6</td>
<td>$t(-1, 1-b+a, 1)$</td>
<td>$x</td>
</tr>
<tr>
<td>7</td>
<td>$t(-1, -1+b-a, -1)$</td>
<td>$x</td>
</tr>
<tr>
<td>8</td>
<td>$t(-1, 1-b+a, 1)$</td>
<td>$x</td>
</tr>
<tr>
<td>9</td>
<td>$t(-1, -1-b+a, -1)$</td>
<td>$x</td>
</tr>
<tr>
<td>10</td>
<td>$t(-1, 1-b+a, 1)$</td>
<td>$x</td>
</tr>
<tr>
<td>11</td>
<td>$t(-1, -1-b+a, -1)$</td>
<td>$x</td>
</tr>
</tbody>
</table>

Table 2: piecewise-linear 2-D maps for $i$

<table>
<thead>
<tr>
<th>$i$</th>
<th>$f_i$</th>
<th>$g_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$\frac{1}{b-a+1} \cdot \frac{(b-1)x_0+1}{b-x_0}$</td>
<td>$\frac{z_0 + (b-1)x_0+1}{b-a+1} - x_0 = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>1</td>
<td>$\frac{2}{b-a+2 \cdot \alpha_i - 2 \cdot \alpha_i - x_0}$</td>
<td>$\frac{b \cdot \alpha_i - 2 \cdot \alpha_i + (b-1)x_0+1}{b-a+2 \cdot \alpha_i} - x_0 = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{(b-1)(5x_0+1)}{b+1} - 1 - \beta$</td>
<td>$\frac{z_0 - (b-1)(5x_0+1)}{b+1} - x_0 + \frac{1}{b-a+1} = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>3</td>
<td>$\frac{(b-1)(x_0+1)}{b-\alpha+1} - z_0 - x_0 - \frac{1}{b-a+1}$</td>
<td>$\frac{z_0 - (b-1)(x_0+1)}{b-\alpha+1} - x_0 + \frac{1}{b-a+1} = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>4</td>
<td>$\frac{(-b+\alpha-1)(5\alpha_i-\alpha_i-x_0+1)}{b} + z_0 - x_0 - \frac{1}{b-\alpha+1}$</td>
<td>$\frac{z_0 - (-b+\alpha-1)(5\alpha_i-\alpha_i-x_0+1)}{b} - x_0 + \frac{1}{b-a+1} = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>5</td>
<td>$\frac{(-b+\alpha-1)(x_0+1)}{b-\alpha+1} - z_0 - x_0 - \frac{1}{b-\alpha+1}$</td>
<td>$\frac{z_0 - (-b+\alpha-1)(x_0+1)}{b-\alpha+1} - x_0 + \frac{1}{b-a+1} = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>6</td>
<td>$\frac{(-b+\alpha-1)(5\alpha_i-\alpha_i-x_0+1)}{b} + z_0 - x_0 - \frac{1}{b-\alpha+1}$</td>
<td>$\frac{z_0 - (-b+\alpha-1)(5\alpha_i-\alpha_i-x_0+1)}{b} - x_0 + \frac{1}{b-a+1} = \frac{1}{b-a+1}$</td>
</tr>
<tr>
<td>7</td>
<td>$\frac{(-b+\alpha-1)(x_0+1)}{b-\alpha+1} - z_0 - x_0 - \frac{1}{b-\alpha+1}$</td>
<td>$\frac{z_0 - (-b+\alpha-1)(x_0+1)}{b-\alpha+1} - x_0 + \frac{1}{b-a+1} = \frac{1}{b-a+1}$</td>
</tr>
</tbody>
</table>
Shishi-odoshi and large deviations
Syuji Miyazaki† and Kenta Yamada‡

†Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan
‡Kyoto-City Horikawa High School, Shibou-Horikawa-Machi 622-2, Nakagyo-ku, Kyoto 604-8254, Japan
†‡ELCAS Kyoto University, http://elcas.kyoto-u.ac.jp/
Email: syuji@acs.i.kyoto-u.ac.jp

Abstract—Shishi-odoshi is a traditional device found in Japanese gardens; composed of a bamboo tube that when filled with water revolves to empty and makes a clanking sound. It consists of a water-filled bamboo tube which clacks against a stone when emptied, and the clack scars beasts and birds from gardens. For a fluctuating flow rate, intervals between the clacks distribute. The flow rate per unit time fluctuates. Although our basic idea was first described in [2] in the context of fluctuations of the flow rate, it can be generalized in a sense that the rate function of a random or chaotic variable can be derived from its first-passage-time problem.

1. Introduction

Shishi-odoshi is a water-filled hydraulic bamboo clapper against a stone when emptied. It is a simple device to drive away birds and animals, which makes a sound by water falling down. Its examples are illustrated in [1]. In this paper, we assume that the flow rate or the amount of water per against a stone when emptied. It is a simple device to drive away birds and animals, which makes a sound by water falling down. Its examples are illustrated in [1]. In this paper, we assume that the flow rate or the amount of water per unit time fluctuates. One may simultaneously a total amount of water at full level. One may regard $f$, $V$ and $n$ respectively as a velocity of a random walker starting from the origin, a distant goal and a first passage time to the goal is reached. Thus, measuring the intervals between the clacks of Shishi-odoshi, we can construct a distribution of the first passage time.

The local average $z$ of the flow rate per unit time is given by

$$z = \frac{\int_{0}^{t_{0}+n} f(t)\,dt}{n} = \frac{V}{n}.$$  

The first passage times $n$ distribute. So do the local averages $z$ due to the above relation. The distribution of $z$ depending on $n$ is denoted as $P(n,z)$, from which we can obtain large deviation statistics of the flow rate per unit time. If $n$ is much larger than its average auto-correlation time of $f(t)$, $P(n,z)$ is scaled as $P(n,z) = P(n,\bar{z}) \exp[-n\phi(z)]$, in which $P(n,\bar{z})$ is an algebraic factor depending on $n$ and $\phi(z)$ is called rate function of the flow rate per unit time [4]. Let $\bar{z}$ be the long-time average as $z$. The rate function is concave up, which satisfies $\phi(z)|_{z=\bar{z}} = \frac{d\phi(z)}{dz}|_{z=\bar{z}} = 0$. As a consequence of the central limit theorem, the rate function is quadratic around $z = \bar{z}$.

In our novel viewpoint inspired by the shishi-odoshi, we observe not directly the local average $z$ or its instantaneous value of the flow rate per unit time but the first passage time $n$ corresponding to the interval between the clacks in the case of shishi-odoshi. The distribution $P(n,z)$ of $z$ can be regarded as a distribution $Q(V,n)$ of $n$ via the relation $z = V/n$.

The transformation of variable from $z$ to $V = nz$ satisfies the conservation of probability $P(n,z)dz = Q(V,n)dV$, so that we have

$$P(n,z) = Q(V,n) \frac{dV}{dz} = nQ(V,n),$$  

$$P(n,\bar{z}) = \bar{n}Q(V,\bar{n}),$$

and the rate function $\phi(z)$ can be indirectly estimated as

$$-\frac{1}{n} \log \frac{nQ(V,n)}{\bar{n}Q(V,\bar{n})}$$ plotted against $z = V/\bar{n}$, where $\bar{n} = V/\bar{z}$ is the long time average of the first passage time.
3. Discussion based on a concrete example

A concrete example is described in the following. An event that a waterdrop falls or does not fall is assumed to occur at regular unit intervals with equal probability, say, according to a fair coin tossing. In this case, $f$ is a binary variable 0 or 1 depending on a integer-valued time step. Note that the water dropping interval in the real dripping faucet is strongly related to each volume of successive waterdrops which may be called the rate flow in this case [5]. The probability $r(n, z)$ that the head appears $nz$ times in $n$ time steps, equivalently the probability that a waterdrop falls $nz$ times in $n$ time steps, yielding the rate flow per unit time $z = \frac{nz}{n}$, is given by

$$r(n, z) = \frac{n^c}{2^n} \frac{(nz)!}{(n - nz)!2^n},$$

which can be expressed by $V$ instead of $z$ as $r(n, V/n) = \frac{n^c}{(V)![(n - V)/2]^n}$. The probability $p(n, V/n)$ that the water container with volume $V$ is filled exactly at time step $n$ is given by

$$p(n, V/n) = \frac{1}{2^n}(n - 1, \frac{V - 1}{n - 1}) = \frac{(n - 1)!}{(V - 1)![(n - V)/2]^n} = q(V, n).$$

Figure 1. $n$-dependencies of this probability are plotted for $V = 2, 3$ and 10. Note that both $P(n, z)$ and $Q(V, n)$ in the preceding section are probability densities and that both $r(n, z)$ and $q(V, n)$ in this section are not probability densities but probabilities.

Taking a large-container limit $V \to \infty$, we apply Stirling’s formula $\log N! \sim N \log N - N$ to the factorials, so that we have the rate function

$$-\frac{1}{n} \log p(n, V/n) = \psi(z) = z \log z + (1 - z) \log(1 - z) + \log 2$$

with $z = \frac{V}{n}$. At the long-time average $z = \bar{z} = 1/2$, the relations

$$\left. \frac{d \psi(z)}{dz} \right|_{z = \bar{z}} = 0$$

are satisfied. In the neighborhood of $z = \bar{z}$, $\psi(z)$ is approximated by the parabola

$$\psi(z) \approx 2 \left( z - \frac{1}{2} \right)^2,$$

which implies the central limiting theorem.

The rate function of the flow rate per unit time can be estimated as

$$-\frac{1}{n} \log \frac{nq(V, n)}{2Vq(V, 2V)}$$

plotted against $z = V/n$, where $\bar{V} = V/\bar{z} = 2V$ is the long time average of the first passage time, which is shown in Fig. 2 for small-container cases $V = 2 (+), 3 (\times)$ and 10 (•) in comparison with the large-container limit (upper line) and the parabola indicating the central limit theorem (lower line). Although the latter holds only around the long time average, it is also drawn outward from the range in application of the central limiting theorem for eye guidance. In spite of small-container cases, a relatively good agreement is observed with the large-container limit. A systematic discrepancy is assumed to come from the fact that the first passage time distribution in a small-container case is asymmetric around the maximum, which eventually becomes symmetric in a large-container limit, as shown in Fig. 1.
4. Concluding remarks

It is difficult to observe fluctuating flow rates in a realistic system, as we mentioned in our preceding study [3].

Our indirect derivation of the rate function from the distribution of the first passage time without observing the instantaneous value $f(t)$ and its local average $z$ can be applied to any stationary fluctuation of $f(t)$, although we confined ourselves to large deviations of the flow rate per unit time inspired by *shishi-odoshi*. One may regard the relation that the sum of a random variable $V$ is equal to the local average $z$ multiplied by the time span $n$ for coarse-graining as the relation that the distance $V$ is equal to the local average of a random velocity $z$ multiplied by the first passage time $n$.

The concrete model described in the preceding section can also be regarded as a one-directional random walk, where the random walker either stops or jumps in a positive direction. A straightforward application to various mathematical or numerical models and experiments describing deterministic chaotic diffusion, where the large deviation statistics are approximately obtained from the first-passage-time distributions.

Acknowledgements

We thank Hiroki Hata for illuminating discussions. Financial support from Kyoto University ELCAS program for high school students is deeply appreciated.

References


Properties of Discrete-Time Bifurcation-Based Amplifiers

Marco Reit, Sven Feldkord and Wolfgang Mathis

Institute of Theoretical Electrical Engineering, Leibniz Universität Hannover
Appelstraße 9A, D-30167 Hannover, Germany
Email: reit@tet.uni-hannover.de, mathis@tet.uni-hannover.de

Abstract—The remarkable filtering characteristics of the mammalian auditory system inspire scientists and engineers to transfer the underlying signal processing principles into technical applications. An established element in the modeling of the nonlinear amplification processes of the hearing organ is the Hopf-type amplifier, that is based on the resonance behavior near the onset of an Andronov-Hopf bifurcation. Following this idea with focus on a digital realization, we show in this contribution first investigations of the nonlinear input-output characteristic of an amplifier based on the Neimark-Sacker bifurcation in the discrete-time domain. We evaluate common features and differences with respect to parameter dependencies.

1. Introduction

One of the main tasks in engineering sciences is the detection and amplification of weak signals. This is the primary processing step for a large number of sensor applications and measurement systems as well as RF- and wireless communication systems. The challenge often lies in the extraction of small desired signals with certain frequencies from a noisy environment. This requires a filtering characteristic with a strong amplification of faint signals in a narrow frequency band. Moreover, a high dynamic range allows the processing of a wide range of signal levels. As a technical example, the lock-in amplifier is a measure-}

acteristic with a strong amplification of faint signals in a narrow frequency band. Moreover, a high dynamic range allows the processing of a wide range of signal levels. As a technical example, the lock-in amplifier is a measure-

In the continuous-time domain, we investigate in this contri-

bution the resonance behavior near the onset of the Andronov-Hopf bifurcation. Several realizations of such an amplifier exist [1, 6–8]. Following the Hopf-type amplifier in the continuous-time domain, we investigate in this contribution the resonance behavior of the equivalent Neimark-Sacker bifurcation in the discrete-time domain. We analyze common features and differences with respect to parameter dependencies. Due to its high dynamic range and its coincident input dependent adaptive bandwidth, this novel digital nonlinear amplifier offers new possibilities in digital signal processing.

2. Hopf-Type Amplifier

The generic Hopf-type amplifier is described by the (truncated) normal form equation of the Andronov-Hopf bifurcation. Using the $\omega_0$-rescaled form [7] and adding the excitation term $a(t)$, the associated differential equation is given by

$$\dot{z} = (\mu + i) \omega_0 z + \sigma \omega_0 |z|^2 z - \omega_0 a, \quad z(t), a(t) \in \mathbb{C}. \tag{1}$$

Here, $i$ is the imaginary unit, $\mu \in \mathbb{R}$ denotes the bifurcation parameter and $\omega_0$ is the natural frequency of oscillation. In general, the coefficient $\sigma$ is a complex quantity $\sigma = \sigma_R + i \sigma_I$. Omitting the excitation and using the substitution $z(t) = r(t) e^{i \phi(t)}$ converts (1) into the polar coordinates

$$\dot{r} = \omega_0 r \left( \mu + \sigma_R r^2 \right), \quad \dot{\phi} = \omega_0 + \sigma_I \omega_0 r^2. \tag{2}$$

Besides the fixed point at the origin, (2) shows the steady-state solution $r = \sqrt{-\mu/\sigma_R}$. Thus, the amplitude of the self-sustained oscillation depends on the real part of the coefficient $\sigma$, in contrast to the oscillation frequency, $\dot{\phi} = \omega_0 - \mu \omega_0 \sigma_I/\sigma_R$, which is affected by both, the real and the imaginary part of $\sigma$. Since the system should serve as an amplifier, it must operate in that region, where the stable fixed point is the only solution. This condition is fulfilled by choosing the parameters to $\sigma_R < 0$ and $\mu \leq 0$. In this
case, the system shows a supercritical Andronov-Hopf bifurcation at $\mu = 0$ [13]. Otherwise, for $\sigma_R > 0$ and $\mu \geq 0$, the single fixed point is unstable and the bifurcation is subcritical [13]. The description in polar coordinates (2) shows the rotational symmetry around the equilibrium point $r = 0$.

This leads to the assumption that the sinusoidal input signal $a(t) = \omega_0 e^{i\omega t}$ in (1) results in the sinusoidal steady-state response $z(t) = z_0 e^{i(\omega t + \phi)}$. Hence, substituting $a(t)$ and $z(t)$ in (1) allows the calculation of the input-output amplitude relation

$$a_0 = \sqrt{\left(\mu z_0 + \sigma_R z_0^2\right)^2 + \left(1 - \omega_0^2 z_0 + \sigma_I z_0^2\right)^2},$$

as well as the phase relation (see [12]). Beside the well studied dependencies of the Hopf-type amplifier response regarding the bifurcation parameter $\mu$ as well as the amplitude $a_0$ and frequency $\omega$ of the excitation [3, 6, 7], recent investigations focus on the nonlinearity coefficient $\sigma$ [9, 11, 12]. The representation of $\sigma$ by the absolute value $\hat{\sigma}$ and the phase $\delta$ as $\sigma = \sigma_R + i \sigma_I = \hat{\sigma} e^{i\delta}$ provides deeper insights in the parameter dependencies of the input-output behavior. The resulting effects on the steady-state response by variation of $\hat{\sigma}$ and $\delta$ are shown in Fig. 1. An increase in $\hat{\sigma}$ leads to a compression of the output amplitude $z_0$. Hence, $\hat{\sigma}$ denotes a damping factor. The variation of $\delta$ leads to a uniform rotation of the output structure around the point $(\omega = \omega_0, \mu = 0)$. In this case, for the interval $\pi/2 < \delta < 3\pi/2$ the Andronov-Hopf bifurcation is of supercritical type. As shown by the dashed line at the cross section $\mu = -0.1$, the variation of $\delta$ causes hysteretic effects in the input-output behavior, that have already been studied for certain relations between $\sigma_R$ and $\sigma_I$ [9, 11]. With these results, the parameter dependent input-output behavior is fully understood.

### 3. Bifurcation-Type Amplifier Realizations

Further investigations of the response behavior of the Hopf-type amplifier to more complicated transient input stimuli first motivates the realization of cochlea specific signal processing principles by means of electronics. Software simulations possess the drawback that solving nonlinear differential equations with high accuracy is time consuming and needs high computing power. Since recently, attention has focused on the nonlinear amplification characteristics of the Hopf-type amplifier, electronic realizations also become attractive for applications. One of those areas refers to hearing aids, where the constraints are on the one hand real-time processing with low latency and on the other hand low power consumption. These conditions also hold for other mobile applications like RF- and wireless communication systems. Another application field are sensors and measurement systems, in which the accuracy and dynamic range are the key parameters. In dependency of the scope of application and its regarded conditions, it must be chosen between an analog, digital or mixed-signal realization. The first analog electronic Hopf cochlea circuit was build up using discrete electronic components, such as operational amplifiers, multipliers, capacitors and resistors to emulate the differential equation by means of analog computing [6, 7]. Shortly after that, a Hopf-type amplifier was build up using a LCR loop with a chain of diodes as the essential nonlinear element [1]. Since this analog circuit represents a kind of van der Pol oscillator, we assume high harmonic distortions based on our investigations [10]. To overcome the drawbacks of an analog realization, for instance the variation and noise of the electronic components, no tunable parameters or only in terms of voltages, external electrical couplings etc., we constructed a first highly flexible and scalable digital realization implemented on a digital signal processor [8]. The main objectives were real-time processing with low latency and low power consumption [8]. Therefore, we choose the explicit 4th-order Runge-Kutta method to compute (1) with a good compromise between accuracy and performance [8]. The digital processing of the Hopf-type amplifier always underlies damping and stability issues depending on the integration method. Thus, we investigate in this contribution the resonance behavior of the Neimark-Sacker bifurcation in the discrete-time domain, which is referred to as the Andronov-Hopf bifurcation for maps [13].

### 4. Neimark-Sacker-Type Amplifier

The normal form of the Neimark-Sacker bifurcation [13] with an added excitation term $p$ is described by the map

$$z \mapsto e^{i\theta} z \left(1 + \gamma + \sigma |z|^2\right) + p, \quad z, p \in \mathbb{C}. \quad (4)$$

Here, $\sigma = \sigma_R + i \sigma_I = \hat{\sigma} e^{i\delta}$ is the nonlinearity coefficient and $\gamma$ is the bifurcation parameter. For a discrete-time
system with an equidistant time interval, the angle $\theta$ can be considered as normalized characteristic frequency $\theta = 2\pi f_0 / f_s$ with the characteristic frequency of the system $f_0$ and the sampling frequency $f_s$. The bifurcation behavior is similar to the Andronov-Hopf bifurcation in the continuous-time domain. Omitting the excitation, the map (4) shows for $\sigma_R < 0$ a supercritical bifurcation at $\gamma = 0$ where a stable closed invariant curve grows for $\gamma > 0$ from a stable fixed point $\gamma \leq 0$ which changes its stability at $\gamma = 0$ [13]. Otherwise, for $\sigma_R > 0$ and $\gamma < 0$ a stable fixed point is surrounded by an unstable closed invariant curve that vanishes by subcritical bifurcation at $\gamma = 0$ where an unstable fixed point remains for $\gamma > 0$ [13]. The bifurcations occur, when the complex-conjugate pair of eigenvalues of (4), which calculates to $\lambda_{1,2} = e^{i\omega} (1 + \gamma)$, crosses the unit circle at $\gamma = 0$ [13]. Moreover, the eigenvalues also crosses the unit circle at $\gamma = -2$. The characteristics of this latter bifurcation become obvious by substituting the translation $\hat{\gamma} = -2 - \gamma$ into (4), that results in

$$z \mapsto e^{i\theta} (1 + \hat{\gamma} - \sigma |z|^2) z + p, \quad z, p \in \mathbb{C},$$

with $\hat{\theta} = \theta \pm \pi (2n + 1) = 2\pi (f_0 \pm (n + 1/2) f_s) / f_s, \quad n \in \mathbb{N}_0$. Thus, the sign of the coefficient $\sigma$ inverts, which causes a change between the supercritical and subcritical bifurcation behavior. Additionally, the characteristic frequency of the bifurcation shifts in terms of the half sampling frequency. In general, since we are dealing with a discrete-time system, the characteristic frequencies of the bifurcations are periodic with the sampling frequency. This refers equally to $\theta$ in (4) which can be written in the more general form $\theta = 2\pi (f_0 \pm n f_s) / f_s, \quad n \in \mathbb{N}_0$. Here, it should be emphasized that for the reconstruction of a continuous-time signal, the Nyquist–Shannon sampling theorem must be hold. Since the map (4) is rotationally symmetric, we assume for $-1 \leq \gamma \leq 0$ that a sinusoidal input signal $p_n = p_0 e^{i\omega t}$ leads to the sinusoidal output signal $z_{n+1} = z_0 e^{i(\omega(n+1) + \sigma_R)}$. Here $n$ denotes the iteration variable. Normalization of the signals on the sampling frequency $\beta = 2\pi \omega / \omega_s = 2\sigma / f_s$ and substituting in (4) allows to calculate the input-output amplitude relation

$$p_0^2 = z_0^2 \left(2 + (\sigma_R^2 + \sigma_I^2) z_0^4 + (2 + \gamma) \gamma + 2\sigma_R (1 + \gamma) z_0^2 - 2 (1 + \gamma + \sigma_R z_0^2) \cos(\beta - \theta) - 2 \sigma_I z_0^2 \sin(\beta - \theta)\right),$$

as well as the phase relation, which is neglected at this point. The algebraic equation (6) describes the amplitude $z_0$ of the output signal caused by a sampled sinusoidal excitation dependent on all given parameters in (4). This allows parametric studies to get deeper insights of the input-output behavior of the Neimark-Sacker-type nonlinear amplifier.

In order to analyze common features and differences against the Hopf-type amplifier in section 2, the resulting output amplitude $z_0$ is plotted over the excitation frequency $f$ and the bifurcation parameter $\gamma$ as illustrated in Fig. 2. Compared to the output amplitude of the Hopf-type amplifier in Fig. 1, huge similarities in the input-output behavior can be noted. Figure 2a) shows that the amplification increases by shifting the bifurcation parameter towards the bifurcation point $\gamma = 0$. Additionally, $\delta$ also denotes a damping factor. Considering a broader range for the excitation frequency $f$ and the bifurcation parameter $\gamma$ in Fig. 2b) points out, that (6) includes the aforementioned periodicity of the resonance structures in terms of the sampling frequency as well as the resonance structure of the subcritical Neimark-Sacker bifurcation at $\gamma = -2$ shifted by $f_s / 2$. Here, it must be noted that the frequency range of discrete-time systems consists of frequencies between $-f_s / 2$ and $f_s / 2$. Thus, analytic signals with a frequency $f$ in the range of $f_s / 2 < f < f_s$ are mapped to analytic signals with the frequency $f = f - f_s$ in the range of negative

\[ \hat{\gamma} = -2 - \gamma \]

\[ z \mapsto e^{i\theta} (1 + \hat{\gamma} - \sigma |z|^2) z + p, \quad z, p \in \mathbb{C}, \]

\[ p_0^2 = z_0^2 \left(2 + (\sigma_R^2 + \sigma_I^2) z_0^4 + (2 + \gamma) \gamma + 2\sigma_R (1 + \gamma) z_0^2 - 2 (1 + \gamma + \sigma_R z_0^2) \cos(\beta - \theta) - 2 \sigma_I z_0^2 \sin(\beta - \theta)\right), \]

\[ \delta = \pi, \quad \sigma = 1 \]

\[ \delta = 3.5 \]
frequencies $-f_s/2 < \tilde{f} < 0$, where the output amplitude and the resonance structure is still the same due to the $f_s$-periodicity. The same applies for frequencies lower than $-f_s/2$ vice versa. To analyze the influence of the imaginary part of $\sigma$ and to compare the behavior with the Hopf-type amplifier in Fig 1, we plot the output amplitude $z_0$ of the Neimark-Sacker system in Fig.3 while using $\sigma_R = 0$ and $\sigma_I = -1$. It is shown, that the behavior strongly differs. Besides the rotation, a bending and connecting of the sub- and supercritical resonance structures occur, which always exists for $\sigma_I \neq 0$. The dashed intersection line at $\mu = -0.25$ discloses, that additionally to the hysteresis behavior, separated solution sets arise. In general, this kind of ambiguities are hard to find by numerical simulations or measurements.

5. Conclusion

In this work, we investigate a nonlinear amplifier that is based on the normal form equation of the Neimark-Sacker bifurcation in the discrete-time domain. Following the idea of the Hopf-type amplifier, we can derive algebraic equations that describe the parameter dependent input-output behavior regarding the properties of the discrete-time domain. This allows parametric studies which disclose ambiguities in form of hysteresis effects and separated solution sets. Since the amplification characteristic shows huge similarities to the Hopf-type amplifier, it could be preferred to a digital realization of the latter one by means of an integration method, which underlies damping and stability issues as well as higher calculation effort. Our analysis gives deeper insights in the input-output behavior of a bifurcation-based amplifier in the discrete-time domain.

References

Detecting directional couplings from time series: joint distribution of distances

Yoshito Hirata† José M. Amigó†,‡, and Kazuyuki Aihara†

†Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
‡Centro de Investigación Operativa, Universidad Miguel Hernández,
Avda. De la Universidad s/n. 03202 Elche, Spain
Email: yoshito@sat.t.u-tokyo.ac.jp

Abstract—When we investigate the dynamics underlying a complex system where we can observe time series of several different components, we want to identify directional couplings between each pair of those components. In such a circumstance, we should not want to assume that the dynamics is linear or that all the components can be observed. Thus, here we propose a method for identifying directional couplings based on a joint distribution of distances. The proposed method will be easily extended to the analysis of point processes.

1. Introduction

There are various networked systems where each component of the systems may be coupled with others. For investigating such networked systems through the observations of some of these components, identifying directional couplings is the first thing we should do. By a directional coupling, we mean the influence that one component exerts upon another. Many methods have been proposed for identifying such directional couplings [1-7]. However, some common drawbacks are that (i) such methods assume the linearity of the underlying systems, (ii) such methods presume that one can observe every component in the networked systems, and (iii) such methods assume a family of models.

To overcome these drawbacks, we proposed in 2010 a method for identifying directional couplings based on time series using recurrence plots [5]. Recently, we proposed two other methods for identifying directional couplings [8]. In addition, we proposed the simultaneous use of the three methods [8]. When evaluating individual performances of the three methods, we found that one of the methods using the joint distribution of distances outperforms the other two methods. Thus, in this presentation, we focus on the joint distribution of distances for a pair of states of components to identify directional couplings.

2. Backgrounds

2.1. Takens’ theorem

Suppose that we are interested in a dynamical system \( f : M \to M \) on a \( m \)-dimensional manifold \( M \) described by \( x_{i+1} = f(x_i) \) for \( i \geq 0 \) given an initial condition \( x_0 \in M \). Let us assume that we can only observe a scalar value \( s_x(i) = g(x_i) \) through an observation function \( g : M \to R \). This poses the problem of how to recover the information of \( x_i \) by using the limited information of \( \{s_x(i) \mid i \geq 0\} \).

To resolve this problem, the key idea is to use delay coordinates, which were proposed by Takens [9]. Given \( \{s_x(i) \mid i \geq 0\} \), delay coordinates corresponding to \( x_i \) can be defined by \( \tilde{s}_x(i) = H_x(x_i) = (s_x(i), s_x(i+1), \ldots, s_x(i+d-1)) \).

Takens [9] showed that if \( d > 2m \), it is a generic property that the relation between \( x_i \) and \( \tilde{s}_x(i) \) is one-to-one. Therefore, in such a case, the following diagram commutes:

\[
\begin{align*}
x_i & \xrightarrow{f} x_{i+1} \\
\downarrow H_x & \quad \downarrow H_x \\
\tilde{s}_x(i) & \xrightarrow{\tilde{f}} \tilde{s}_x(i+1)
\end{align*}
\]

By using the delay coordinates, we can reconstruct a dynamical system that is equivalent to the original dynamics \( f \). Thus, even if we can neither observe \( x_i \) nor know the original dynamics \( f \), we can learn the underlying dynamics and even predict the future of \( s_x \) by \( \tilde{f} \).

2.2. Stark’s theorem

The above theorem by Takens has been extended by Stark to the forced system [10]. Mathematically, a setting for a forced system can be written as follows: We have two dynamical systems \( f : M \to M \) and \( g : M \times N \to N \), where \( M \) and \( N \) are manifolds of \( m \) and \( n \) dimensions, respectively. Therefore, the first system \( f \) is an autonomous system, while the second system \( g \) is forced by the input of the first system. In what follows, we use the following notations for...
describing states for these systems: 

\[ x_{i+1} = f(x_i), \quad y_{i+1} = g(x_i, y_i) \quad \text{for} \quad i \geq 0. \]

Suppose that we do not have access to the first system and we can observe a scalar value \( s_y(i) = h_y(y_i) \) depending only on the state for the second system. Similarly, we construct delay coordinates by \( \tilde{s}_y(i) = H_y(x_i, y_j) = (s_y(i), s_y(i+d-1) \ldots , s_y(i+1), j) \).

Then it is a generic property that if \( d > 2(m+n) \), the joint set of \( (x_i, y_j) \) and \( \tilde{s}_y(i) \) are one-to-one on the attractor [10]. This theorem means that in the forced system, we can reconstruct the information of not only the forced system but also the driving force.

### 2.3. A method using recurrence plots for inferring directional couplings

The first person who used the above Stark’s theorem in an application is Timothy D. Sauer [11]. He assumed that one can observe several forced systems and proposed how to reconstruct the common driving force.

The second application of Stark’s theorem is by two of us for inferring directional couplings [5]. We used an implication of Stark’s theorem for denying the existence of a directional coupling. Let us compare two sets of delay coordinates \( \tilde{s}_y(i) \) and \( \tilde{s}_y(i) \). When the system of \( f \) drives the system of \( g \) as defined as above and \( d \) is sufficiently large, \( \tilde{s}_y(i) \) is one-to-one with \( x_i \), while \( \tilde{s}_y(i) \) is one-to-one with \( (x_i, y_i) \). Hence, if two sets of delay coordinates \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are close to each other, \( (x_i, y_j) \) and \( (x_j, y_j) \) are close to each other. This relation means that \( x_i \) and \( x_j \) are close to each other, implying that \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are close to each other.

To infer the directional coupling, Ref. [5] used the contraposition of the above relation: Namely, if \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are close to each other while \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are not close to each other, then we can deny a directional coupling from a system described by \( x \) to a system described by \( y \). In Ref. [5], this statement is tested by using recurrence plots [12].

### 3. The proposed method

#### 3.1. Joint distribution of distances for identifying directional couplings

Similarly to Ref. [5], we used Stark’s embedding theorem for inferring directional couplings [8]. This time, we interpret the theorem more straightforward.

When we use Stark’s embedding theorem, if the system described by \( x \) drives the system described by \( y \), and \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are close to each other, in other words, \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are neighbors, then \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) are also neighbors. Therefore, if we evaluate the similarity between \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \), and the similarity between \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) by the corresponding distances, these distances can be plotted in a two-dimensional space as shown in Fig. 1 and can occupy the triangle region as shown by red in Fig. 1. This is the idea we would like to use here for inferring a directional coupling from a system to another.

![Fig. 1. The distance between \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \) and the distance between \( \tilde{s}_y(i) \) and \( \tilde{s}_y(j) \).](image-url)

#### 3.2. Implementation

In Ref. [8], we proposed the following implementation for inferring directional couplings using the above idea: First, we calculate the Euclidean distances between every pair of embedded state vectors for the system described by \( x \) and those for the system described by \( y \). Second, we convert the distributions of distances so that the distances may not be able to call them as distances strictly anymore. Third, we divide the axis of the distances for \( x \) uniformly into \( 2B \) bins. For each bin \( b \), we find the minimum distance \( D_y(b) \) for \( y \). Lastly, we evaluate the distribution

\[ \Delta = \left\{ D_y(B + b) - D_y(b) \right\} | b = 1, 2, \ldots, B \]

Therefore, this distribution is expected to be biased towards the positive side if there is a directional coupling from \( x \) to \( y \) and the joint distribution of distances looks like one in Fig. 1. Thus, we construct a one-sided t-test based on the null hypothesis that the mean of \( \Delta \) is zero. If this distribution can reject the null hypothesis, then we declare that there is a directional coupling from the system described by \( x \) to the system described by \( y \).

### 4. Examples
In Ref. [8], we compared the above method of joint distribution of distances with other methods. But, here we only show the results of the above method because the space of this proceedings is limited. For detailed comparisons of the methods, see Ref. [8]. The p-values shown below are the p-values obtained by the above t-test discussed in Section 3.2.

Fig. 2. Coupling configurations considered in Sections 4.1 (A), 4.2 (B) and 4.3 (C).

4.1. Mutually coupled logistic maps

First, we tested the above method using mutually coupled logistic maps (Fig. 2A). We varied the coupling strengths between 0 and 0.2, and generated a time series of length 1000 for each pair of coupling strengths.

Fig. 3. Results for tests for inferring directional couplings given time series generated from mutually coupled logistic maps. In each panel, the gray scale shows the logarithm of the p-value with base 10.

Fig. 4. Example for joint distribution of distances. Here we used a logistic map (driver) unidirectionally driving another (driven system). This example corresponds to $\eta_{xy} = 0$ and $\eta_{yx} = 0.15$ of Fig. 3.

The results are summarized in Fig. 3. The proposed method could properly infer the directional couplings if the coupling strength is modest or stronger, and the opposite coupling strength is not too strong. In addition, we show an example of joint distribution of distances in Fig. 4.

4.2. Logistic maps driven by another

Second, we tested the proposed method using logistic maps driven by another logistic map (See Fig. 2B). Here we assume that the two logistic maps we can observe are not coupled. The other simulation conditions are similar to the first example.

The results are shown in Fig. 5. We can see that the proposed method was not influenced by the common hidden driver. Thus, it seems that the proposed method does not induce false positive results, under which we declare, due to the influence of the common hidden driver, that there exists directional couplings between observed systems. This point means that the proposed method has a nice property that most of the existing methods do not have.

Fig. 5. Results of tests for inferring directional couplings given time series generated from two logistic maps driven by another logistic map. Here, the two logistic maps we can observe do not have direct couplings. See the caption of Fig. 3 for the interpretations.

4.3. Mutually coupled logistic maps driven by another

Third, we attempt to infer directional couplings under the existence of a common hidden driver (see Fig. 2C). Here, we also use logistic maps. The other simulation conditions are similar to the previous cases.

Fig. 6. Results for tests for inferring directional couplings given time series generated from mutually coupled logistic maps driven by another. See the caption of Fig. 3 for the interpretations.
The results are shown in Fig. 6. These results look similar to Fig. 3. Namely, the proposed method can work well and infer directional couplings even if there is a common hidden driver.

5. Discussions

We have proposed a method for inferring a directional coupling based on time series. Our method uses Stark’s embedding theorem and infers a directional coupling using the joint distribution of distances. The method seems to work properly even if there exists a common hidden driver. Therefore, we do not have to be able to observe the entire components of a network system when we investigate its topology.

In Ref. [8], we also demonstrated that the proposed method works well even if the length of time series is 250, showing examples based on real data. The strength of the proposed method is that we use distances. Therefore, in Ref. [8], we applied our method for irregularly sampled data. Being able to obtain distances is a common condition for nonlinear time series analysis of exotic data such as a time series of network [13] and marked point processes [14-17]. Therefore, we believe that by combining with other methods, the proposed method will help us to study multivariate time series data.

Acknowledgments

This research was supported by the Platform Project for Supporting Drug Discovery and Life Science Research (Platform for Dynamic Approaches to Living Systems) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) and the Japan Agency for Medical Research and Development (AMED). This research is also supported by JSPS KAKENHI Grant number 15H05707. In addition, we thank the three anonymous reviewers for their constructive comments. Especially, the second reviewer suggested to use a regression analysis instead of the implementation discussed in Section 3.2. That could be one of interesting implementations.

References

On ranking and mining universities with the encyclopedia Wikipedia

Zongjian Li†, Cong Li†‡, Xiang Li†‡
†Adaptive Networks and Control Lab, Department of Electronic Engineering, Fudan University, Shanghai 200433, P.R.China
‡Research Center of Smart Networks and Systems, School of Information Science Engineering, Fudan University, Shanghai 200433, P.R.China
Email: {zjli13, cong_li, lix}@fudan.edu.cn

Abstract—The large data of Wikipedia has motivated new research branches, such as the evaluation of the reputation of an entry in its field by utilizing the Wikipedia data, where we can dig the relationship of entries via the analysis of Wikipedia data. In this work, we extensively evaluate the university entries in the Wikipedia to rank the universities over the whole world. Several reputation indicators for Wikipedia entries are introduced and compared with the QS and THE university ranking, while the in-degree reputation indicator has a strong correlation with QS and THE ranking. We propose two data mining methods to generate effective Wikipedia article reference subnetworks and we find that the community property of the university article reference subnetworks can reflect the geographic distribution of the universities.

1. Introduction

Wikipedia is a free-access and web-based multilingual encyclopedia that voluntaries from all around the world can write and edit. The English Wikipedia with more than 5 million articles is the largest one among the 291 Wikipedia editions, and has become one of the most popular public collaborative information repository [1]. Previous study on Wikipedia most focus on its collaborative systems, i.e. edit patterns [2–5]. Besides the relationship network of editors in Wikipedia, the articles of Wikipedia form a large-scale complex networks, the Wikipedia article reference networks (WARN). Articles are regarded as nodes, which are connected by the URL links. The trustreputation management system [5] of Wikipedia guarantees that the article jumps are based on the reputation of articles, in other words, articles with more links are high-reputation articles.

The massive data from Wikipedia provides us with a new resource to dig the relationship between things. One interesting question is whether the reputation of an article represents the rank of the entry of the article in its field. For instance, does the article of “google” entry in Wikipedia have more links than that of the article of an unnameable company? In this paper, we study this question by taking the rank of universities as the research subjects, since the rank of universities has been widely discussed in our daily life, and are associated to people’s life. We analyze the relationship between the reputation of the university articles in Wikipedia and the human-defined university ranking, including the World University Rankings by Quacquarelli Symonds (QS) and by Times Higher Education (THE) magazine. The list of 114 chosen universities includes the overlapping universities of the top 100 in QS ranking and the top 200 in THE ranking, as well as the overlapping universities of the top 100 in THE ranking and the top 200 in QS ranking. All the data of Wikipedia, QS ranking and THE ranking are collected by 2015.

This paper is organized as follows. In Section 2, we study the relation between the reputation indicators for Wikipedia entries with the QS and THE rankings. In Section 3, we propose two methods to mine the effective Wikipedia article reference subnetworks (EWARS). In Section 4, we apply the EWARS to investigate the relationship between universities. Finally, we conclude in Section 5.

2. Reputation indicators for Wikipedia entries

The reputation of the entries can be evaluated by different indicators, such as the length of an article and the number of the revisions of the article. Three types of indicators, the intuitive criterion, the potential criterion and the deep seated criterion are discussed in this work. The intuitive criterion can be obtained directly when you read an article, for instance, the length of an article which is counted in bytes. The potential criterion is a kind of previous record of the article, such as the number of the revisions of the article, the number of editors who have rewritten the article, and the time of the edits of the article in one year. The relation between these two types of criterions and the QS or THE ranking are shown in Figure 1. The “T” and “t” (“Q” and “q”) marks denote that their x-axis is the THE (QS) ranking. Marks in upper-case letter mean that the universities locate in an English-speaking region, and the lower-case letter marks represent the universities in a non-English-speaking region. The 114 universities are located in 20 countries and regions, and 6 of them are English-speaking countries. Totally, 72 universities are in the English-speaking region.

The deep-seated criterion takes into account not only the inherent properties of an article but also the relationship between the article and others. Articles are regarded as nodes and the URL links as connections between articles. This study only adapts articles in English Wikipedia, while the external links to other web sites, such as non-
The number of revisions

<table>
<thead>
<tr>
<th>University</th>
<th>Degree</th>
<th>Out Degree</th>
<th>In Degree</th>
<th>Degree vs. Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>University 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>University 2</td>
<td>0.1</td>
<td>10</td>
<td>6</td>
<td>0.2</td>
</tr>
<tr>
<td>Other 1</td>
<td>3000</td>
<td>12000</td>
<td>8000</td>
<td>4</td>
</tr>
<tr>
<td>Other 2</td>
<td>1000</td>
<td>6000</td>
<td>3000</td>
<td>2</td>
</tr>
<tr>
<td>Other 3</td>
<td>8000</td>
<td>12000</td>
<td>6000</td>
<td>3</td>
</tr>
<tr>
<td>Other 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other 5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1: Relation between the intuitive reputation criterion or the potential reputation criterion of universities in Wikipedia and their QS or THE ranking.

English Wikipedia sites, non-article page and self-loop site are not counted in. The Wikipedia article reference networks (WARN) are directed networks (see Figure 2). Outgoing links direct to another article of a key-word that appears in current article, however, the sources of in-coming links can be any article in Wikipedia. Figure 3 shows that the QS or THE ranking is more strongly linear correlated with the in-degree reputation indicator than with the out-degree reputation. Inspired by [6], the sum of the degree of two-hopcount neighbors in WARN is calculated and compared with the QS or THE ranking (See Figure 4). Same with existing research [7], we find that the WARN has small-world properties, i.e. small average shortest path and large clustering coefficient [8]. This phenomenon implies that a relative large hopcount of a node may cover most part of the whole network. Hence, we only consider the sum of in-degree and out-degree of the node itself and its 1-hopcount or 2-hopcount neighbors.

The linear correlation coefficients between three types of criterions and the QS or THE ranking are shown in Figure 5. For the universities located in English-speaking countries, the deep-seated reputation criterion performs better than other types of criterions in characterizing the university ranking.
3. Two data mining methods to generate effective Wikipedia article reference subnetworks

The Wikipedia article reference network (WARN) involves all articles in Wikipedia, however, only a small part of the articles is essential and interesting for researchers from different fields. It is a challenge to extract the effective Wikipedia article reference subnetworks (EWARS), which contain only the needed articles. In this work, we design two methods to generate the EWARS, and take the EWARS of the 114 universities as an example.

3.1. Path length (PL) Method

First, we generate a pre-EWARS, which is composed of the articles of the 114 universities along with their 1-hopcount neighbor articles connected by the in-coming links (or out-going links), and all the connections between these articles. Second, we convert this directed network to an undirected pre-EWARS by adding inverse links. Note that the WARN is a directed network which contains incoming links and out-going links. The study in Section 2 has shown that the neighbours connected by in-coming links and the out-going links of the universities are considerably different, thus, there are two types of pre-EWARS. Third, the shortest path length between any two university articles is calculated, and the length represents the strength of the correlation between two universities. Then, the shortest path length threshold will be set to obtain the EWARS. Two university articles with an equal or smaller shortest path length than the threshold will be connected by an undirected link, otherwise disconnected.

3.2. Vertex connectivity (VC) Method

Similarly to PL Method, we first generate the pre-EWARS, and then remove connections between university articles and connections between non-university articles. In other words, only undirected connections between a university article and a non-university article are kept (see Figure 6). Next, the vertex connectivity between any two university articles are calculated to characterize the correlation between two universities. The vertex connectivity is defined as the smallest number of nodes to remove that makes no path between two nodes. Finally, a vertex connectivity threshold will be set to generate the EWARS. Two university articles with a vertex connectivity over the threshold will be connected by an undirected link, otherwise disconnected.

4. Application of the EWARS

We obtain an undirected pre-EWARS of the 114 university articles with their out-going connected articles. There are 29,416 nodes and 1,898,348 undirected links in this pre-EWARS. With the EWARS generated by using PL method, there are 1,900 pairs of university articles have shortest path length of 1, while 4,385 pairs of university articles have shortest path length of 2, and 156 pairs of university articles have shortest path length of 3. That means each pair of the 114 university articles has a path length shorter than 4, so this pre-EWARS is a strongly-connected network. When we set a threshold of the shortest path length as 1, we get an EWARS of the university articles (see Figure 7). The (red) nodes in the middle are the largest clique [9], which has 40 universities including 37 American universities and 2 Canadian universities. However, we find the EWARS is already a dense network even when we choose the smallest threshold of 1. The detail relationships between the universities are difficult to be digged.

We use the VC method to generate the EWARS of the 114 university articles with their out-going connected articles. Compared to the shortest path, the vertex connectivity between any two articles has a large range of values, and approximately follows binomial distribution (see Figure 8). The quartiles of the vertex connectivity thresholds are $T = 69, 91, 117$. EWARS with a threshold $T = 172$ is shown in Figures 9. In Figure 9, university articles are automatically divided in groups by their relation with other university articles by cartographic software, so the groupings are just for demonstration. Specially, links in the largest and the second largest clique are colored as red and yellow, respectively. The left part consists of universities in America, two red dots in the middle are universities in Canada, the sphere at the bottom consists of universities in UK, the right-top part is mainly occupied by universities in Hong Kong, China, Australia and Singapore. Notice that the results for the EWARS of the 114 university articles with their in-coming connected articles has similar results with the above discussion. The results demonstrate that the community property of the EWARS generated by the VC method matches the geographic distribution of the universities. It implies that the relationship between entries in Wikipedia can be reflected by the properties in EWARS.

5. Conclusion

In this work, we study several reputation indicators for entries in Wikipedia. We take the university articles as an example, and compare them with the QS and THE university rankings using internal and external properties of articles. The linear correlation coefficient between the indicators and the QS and THE rankings are also calculated. We find that the in-degree reputation criterion has the most strong correlation with the QS or THE rankings for the u-
universities located in English-speaking countries. We then propose two data mining methods, based on the shortest path length and the vertex connectivity, to generate effective Wikipedia article reference subnetworks. An interesting finding is that the community property of the effective university article reference subnetworks well matches the geographic distribution of the universities.

Acknowledgments

This work was partly supported by the Natural Science Fund for Distinguished Young Scholar of China (No. 61425019), the National Natural Science Foundation (No. 61273223), and Natural Science Foundation of Shanghai (No. 16ZR1444400).

References


Figure 7: The EWARS of the university articles generated by PL method with threshold $T = 1$.

Figure 8: The distribution of vertex connectivities between any two articles in pre-EWARS generated by VC method.

Figure 9: The EWARS obtained by utilizing the VC method with a vertex connectivity threshold $T = 172$.


Characterization and control of chimera state collapses.

Ralph G. Andrzejak†, Christian Rummel*, Florian Mormann§ and Kaspar Schindler‡

†Universitat Pompeu Fabra, Department of Information and Communication Technologies, Barcelona, Spain
*Support Center for Advanced Neuroimaging, University Institute for Diagnostic and Intervventional Neuroradiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland
§Department of Epileptology, University of Bonn, Bonn, Germany
‡Schlaf-Wach-Epilepsie-Zentrum (SWEZ), Department of Neurology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland

Email: ralph.andrzejak@upf.edu

Abstract– Studying signals from a mathematical model of a statistical physics dynamics, we show that an outbreak of global synchronization in a partly synchronized network is promoted by a synchronization drop in the weakly synchronized network part, rather than critically high synchronization [1]. This strikingly counterintuitive mechanism can be found also in nature, as we exemplify by epileptic seizures, indicating relevance for neurology and neuroscience. Our control scheme that applies this counterintuitive mechanism succeeds in both provoking and preventing global synchronization outbreaks. Further potential applications of this scheme include power-grid stabilization in electrical engineering and the therapy of various neurological diseases.

Acknowledgments

R.G.A. and F.M. acknowledge funding from the Volkswagen foundation. R.G.A was supported by the Spanish Ministry of Economy and Competitiveness (Grant FIS2014-54177-R). This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 642563 (R.G.A.). F.M. acknowledges support from the German Research Council (DFG, MO 930/4-1 and SFB 1089) and the European Commission (FP7 602102 EPITARGET). K.S. and C.R. are grateful for support by the Swiss National Science Foundation (Projects No. SNF 320030_122010, No. 33CM30-140332, No. 32003B_155950).

References

Extremal optimisation approaches for building complex geometric structures in 3D integrated circuits

Maciej J. Ogorzalek†, Katarzyna Grzesiak-Kopec† and Piotr Oramus†

†Department of Information Technologies
Faculty of Physics, Astronomy and Applied Computer Science
Jagiellonian University
ul. Łojasiewicza 1, 30-348 Krakow, Poland
Email: maciej.ogorzalek@uj.edu.pl

Abstract—The 3D integrated circuit design task poses an extremely difficult intellectual challenge. Modern integrated circuits are composed of thousands of building blocks and billions of transistors. Also an additional challenge is posed by realisation of various blocks using heterogeneous technologies such as CMOS of various technology nodes, batteries and super capacitors, sensors, RRAM memories and many others. Solution of the extremely difficult problem of building block geometric positioning and their interconnects. The solution has to meet a number of specific requirements and satisfy a variety of constraints. Efficient search of huge and discontinuous solution spaces requires new non-deterministic and heuristic algorithms. The goal of our research is to minimize the total wire-length of interconnects between sub-circuits. The paper presents a knowledge intensive 3D ICs layout hypergraph representation together with the elaborated neighborhood optimization heuristics. The results of the Extremal Optimization (EO) implementation applied to the MCNC set of benchmark circuits are reported.

Xiao Fan Liu†

†School of Computer Science and Engineering, Southeast University, Nanjing, China
Email: xfliu@seu.edu.cn

Abstract—Professional association football is a game of talent. The success of a professional club hinges largely on its ability of assembling the best team. Building on a dataset of player transfer records among more than 400 clubs in 24 world-wide top class leagues from 2011 to 2015, this study aims to relate a club’s success to its activities in the player transfer market from a network perspective. We confirm that modern professional football is indeed a money game, in which larger investment spent on the acquisition of talented players generally yields better team performance. However, further investigation shows that professional football clubs can actually play different strategies in surviving or even excelling this game, and the success of strategies is strongly associated to their network properties in the football player transfer network.

1. Introduction

Association football, also referred to as soccer, is probably the sport that gained most global popularity since the 20th century. Professional football industry has thereby grown into a prosperous playground for wealthy investors and big name companies. It was reported that the combined revenue of the top 20 earning clubs in season 2014/15 was over 6.6 billion euros [1]. Meanwhile, the abundant football statistics gathered have attracted fans and scholars to analyze this fascinating game in a quantitative way [2]. However, the achievement of a club is ultimately decided by a squad of 30 players who play the matches through the entire season. Therefore, it is eventually the club’s ability of gathering most talented players which decides about the outcome of the billion euros’ investments.

Actually, talent is believed to be the most important asset in any organization. From senior executives in public companies to common laborers in a massive amount, the acquisition of talents is found to be significantly related to the accomplishment of companies or even national economies [3]. A professional football club usually acquires players from other clubs either permanently, i.e., transfer, by exchanging certain compensation, i.e., transfer fee. The football player movement has already drawn attention from the academia since decades [4]. However, few studies have addressed the relationship between the activities in the transfer market and the success of a club in a systematic way. In this paper, we try to solve this problem by employing a network perspective analysis to the global football player transfer market.

Many natural and man-made systems composed of connected components can be modeled by networks, whose properties are proven to be associated with the functionalities of the systems or their components. Scale-free structure is a common macroscopic property found in social networks. It has been shown that this property could be the cause of the unpredictability of epidemic spreading [5]. Microscopic network such as the topological properties of individual nodes and edges are also widely used to measure the functionalities of system components. They were successfully used to characterize the importance of web pages [6] and to identify influential spreaders in social networks [7].

Building on a dataset of transfer records from 2011 to 2015 of 410 professional clubs in 24 world-wide top class leagues, our work analyzes the properties of the global football player transfer network at both macroscopic and microscopic scales. In this network, nodes are the elite clubs, and the directed edges connecting the nodes are the player transfers. Particularly, the relationship between node properties and the functionalities of professional clubs is studied. Our results show that clubs’ match performance and profitability from the transfer market are strongly associated with the coreness and brokerage properties of their corresponding nodes in the player transfer network.

2. Results

2.1. Construction of the transfer network

In the 5 year period from 2011 to 2015, there were totally 8948 transfer actions among the elite clubs, constituting a player transfer network of 410 nodes connected by 6316 directed edges, where the direction of edges denotes the direction of player movements. All the nodes in the network are strongly connected except for three nodes with only outgoing edges and no incoming edges. The mean shortest distance between all nodes is 2.8 and the clustering coefficient is 0.21. The clustering coefficient is the average ratio between all directed triangles actually formed by each node and the number of all possible triangles that the node could form [8]. Comparing to the average mean shortest distance...
of 2.5 and the average clustering coefficient of 0.04 in random networks of the same size and connection density, the player transfer network exhibits small-world phenomenon. Figure 1A shows the distributions of in-degree $k_{in}$ and out-degree $k_{out}$ of the network. Figure 1B and C show the correlation of in-degree and out-degree and the distribution of excess degree $k_{ex} = k_{in} - k_{out}$ of all nodes respectively. It is shown that the numbers of clubs that a professional club “buys from” and “sells to” are basically equal.

2.2. Club functionalities versus network properties

In this section, we will explore the relationship between the functionalities of a club and its network properties. The ultimate measure of the success of a commercial organization is its profitability, which also applies to a professional football club. Generally speaking, clubs with the highest achievements in prestigious competitions are also the ones that generate the largest revenue from various commercial activities [1]. Meanwhile, a club could also profit directly from the transfer market, by receiving more compensation from the players transferred out than it pays to acquire new players. Therefore, the club functionalities can be described either by its match performance or its transfer profit. Match performance includes the domestic and international match results. We quantify domestic performance of a club by the average game points in its domestic league matches from 2011 to 2015. On the other hand, the five year aggregate IFFHS Club World Ranking (CWR) point is employed to quantify the overall performance of a club in both domestic and international competition [9]. The ability of profiting from player transfers are defined by two measures, i.e., the average annual transfer balance and the cumulative price overflow from player transfers. If a player has transferred from club A to club B then to club C, then we define that the player has transited through club B. The price overflow of this player in club B is the difference between the transfer fees paid by club C to club B and by club B to club A. The correlations between match performance and the transfer profitability are very low, so that they can be considered independent indicators of the functionalities of a professional football club.

Which factors of the transfer network properties affect the functionalities of professional football clubs? The node properties of concern include two categories of metrics that measure the coreness and brokerage of the nodes, respectively. The coreness of a node measures the richness of its connections and is usually described by the number of direct connections combined with the number of indirect connections. The coreness metrics could indicate the search breadth of the club scouts in the transfer market. In this paper, eigenvector centrality and PageRank centrality are chosen as the coreness metrics. The brokerage of a node measures the extent to which it controls the network flow. Clubs with large brokerage metrics tend to exclusively control certain transfer resources and act as brokers amongst other clubs. In this paper, effective size [10], closeness centrality and betweenness centrality are chosen as the brokerage metrics. Table 1 shows Kendall’s Tau between club functionalities and network properties. The match performance, especially international performance, of the clubs is positively correlated with the brokerage metrics, while clubs’ profitability from transfers is generally weakly or not correlated with the network properties. This result suggests that the brokerage power of a club plays the most significant role in determining its match outcome, while the relationships between other network measures and club functionalities are relatively weak.

2.3. International versus domestic transfer networks

International sports labor migration shows different characteristics from domestic sports labor movements [11]. In this paper, the football player transfer network can also be separated into two subnetworks accordingly. The domestic transfer network contains only transfers within a same league and the international transfer network contains only international transfers between different leagues. Table 2 shows the correlation of network properties and club functionalities in both international and domestic transfer networks. It is shown that the node properties taking account of global network connections, i.e., eigenvector centrality, PageRank centrality, betweenness centrality and closeness centrality, are better indicators of both domestic and international match performance, while the node property taking account of local network connections, i.e., effective size, in the domestic transfer network is a better indicator of match performance.

2.4. Money leagues and farm leagues

Although that the match performance and profitability of clubs are generally weakly or not related, a closer examination shows that the clubs in different leagues exhibit different characteristics in the relationship between the two kinds of club functionalities. In some leagues, the more clubs spend, the better match performance they achieve; the leagues with positive correlation and adjusted $R^2 > 0.25$ are referred to as “money leagues”, as the more the clubs spend, the better match performance they achieve; the leagues with positive correlation and adjusted $R^2 > 0.25$ are referred to as “farm leagues”, as the more the clubs profit from the transfer market, the better match performance they achieve. In money leagues, the clubs’ performance are strongly related to their abilities in raising transfer fund. In farm leagues, the clubs’ performances are strongly related to their abilities of profiting from player transfer.

The relationship between the node properties in the player transfer network and the functionality of clubs in
Figure 1: Degree distributions of the global football player transfer network. A: The in-degree and out-degree distributions of the network. B: The in-degree/out-degree relation for each club in the transfer network. C: The distribution of excessive degree $k_{EX}$ of all nodes in the transfer network. The standard deviation of $k_{EX}$ is 5.7.

Table 1: Kendall’s Tau between network properties and club functionalities in the player transfer network.

<table>
<thead>
<tr>
<th></th>
<th>Eigenvector centrality</th>
<th>PageRank centrality</th>
<th>Effective size centrality</th>
<th>Betweenness centrality</th>
<th>Closeness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. league pts</td>
<td>0.18</td>
<td>0.12</td>
<td>0.34</td>
<td>0.34</td>
<td>0.33</td>
</tr>
<tr>
<td>Agg. CWR pts</td>
<td>0.36</td>
<td>0.14</td>
<td>0.48</td>
<td>0.37</td>
<td>0.49</td>
</tr>
<tr>
<td>Balance</td>
<td>0.04</td>
<td>-0.02</td>
<td>0.19</td>
<td>0.16</td>
<td>0.26</td>
</tr>
<tr>
<td>Price overflow</td>
<td>0.04</td>
<td>0.04</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 2: Kendall’s Tau between network properties and club functionalities in international and domestic transfer networks.

<table>
<thead>
<tr>
<th></th>
<th>Eigenvector centrality</th>
<th>PageRank centrality</th>
<th>Effective size centrality</th>
<th>Betweenness centrality</th>
<th>Closeness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. league pts</td>
<td>International</td>
<td>0.29</td>
<td>0.19</td>
<td>0.03</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>Domestic</td>
<td>-0.04</td>
<td>-0.06</td>
<td>0.34</td>
<td>0.04</td>
</tr>
<tr>
<td>Agg. CWR pts</td>
<td>International</td>
<td>0.40</td>
<td>0.24</td>
<td>0.12</td>
<td>0.36</td>
</tr>
<tr>
<td></td>
<td>Domestic</td>
<td>0.13</td>
<td>-0.05</td>
<td>0.45</td>
<td>0.10</td>
</tr>
<tr>
<td>Balance</td>
<td>International</td>
<td>0.05</td>
<td>0.04</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Domestic</td>
<td>0.07</td>
<td>0.04</td>
<td>0.16</td>
<td>0.10</td>
</tr>
<tr>
<td>Price overflow</td>
<td>International</td>
<td>0.03</td>
<td>0.03</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Domestic</td>
<td>0.02</td>
<td>0.04</td>
<td>0.06</td>
<td>0.01</td>
</tr>
</tbody>
</table>

different league categories is shown in Table 3. The parameter correlations of clubs in “money leagues” generally agree with the average nodes in the transfer network. However, in “farm leagues”, the correlation between eigenvector centrality, effective size, closeness centrality and club domestic match performance vanishes, but strong correlation emerges between node coreness and brokerage and club profitability. Especially, the brokerage of a club in the player network is strongly correlated to its annual balance. This phenomenon suggests that clubs in different financial environments actually have various strategies of achieving success, either by acquiring the best players at all costs, or by cultivating players with potential and profit from the re-selling of these valuable assets. More importantly, the successes of both strategies are strongly related to the clubs’ network properties in the global player transfer network.

3. Discussion

Football is probably the most popular sports in the world. The abundance of statistics regarding teams’ activities on and off the pitch has attracted extensive quantitative analysis by fans and scholars from various perspectives. However, despite of the importance of the acquisition of talented players to the success of a professional football club, previous studies rarely addressed the relationship of the clubs’ functionality to their activities in the player transfer market. To do so, we have collected exhaustive transfer records among more than 400 football clubs in major professional leagues from different countries during the last 5 years. Data reveals that football is indeed a money game, in which clubs spend large amounts of money on football stars in order to achieve prestigious status and generate commercial revenue. However, in this winner-takes-all game, the finan-
Table 3: Kendall’s Tau between network properties and club functionalities in different league categories.

<table>
<thead>
<tr>
<th></th>
<th>Eigenvector centrality</th>
<th>PageRank centrality</th>
<th>Effective size</th>
<th>Betweenness centrality</th>
<th>Closeness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. league pts.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Money Leagues</td>
<td>0.25</td>
<td>0.15</td>
<td>0.33</td>
<td>0.29</td>
<td>0.34</td>
</tr>
<tr>
<td>Farm Leagues</td>
<td>-0.01</td>
<td>0.03</td>
<td>0.06</td>
<td>0.15</td>
<td>0.04</td>
</tr>
<tr>
<td>Agg. CWR pts.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Money Leagues</td>
<td>0.30</td>
<td>0.08</td>
<td>0.44</td>
<td>0.39</td>
<td>0.51</td>
</tr>
<tr>
<td>Farm Leagues</td>
<td>0.43</td>
<td>0.25</td>
<td>0.53</td>
<td>0.41</td>
<td>0.51</td>
</tr>
<tr>
<td>Balance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Money Leagues</td>
<td>-0.13</td>
<td>-0.14</td>
<td>-0.09</td>
<td>-0.09</td>
<td>-0.03</td>
</tr>
<tr>
<td>Farm Leagues</td>
<td>0.35</td>
<td>0.20</td>
<td>0.56</td>
<td>0.46</td>
<td>0.59</td>
</tr>
<tr>
<td>Price overflow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Money Leagues</td>
<td>0.03</td>
<td>-0.01</td>
<td>-0.04</td>
<td>-0.05</td>
<td>-0.03</td>
</tr>
<tr>
<td>Farm Leagues</td>
<td>0.19</td>
<td>0.14</td>
<td>0.23</td>
<td>0.17</td>
<td>0.23</td>
</tr>
</tbody>
</table>

cial abilities of clubs are severely unequal. Wealthy clubs with overwhelming financial resources could spend tens or even thousands times more money than normal clubs on acquiring better players, therefore other clubs must seek different strategies to survive in this competitive industry.

Network science provides a systematic perspective and a variety of tools to quantitatively study the structure of complex systems. Particularly, the network properties of system components are found closely related to their functionalities. In this work, we have employed a network perspective of analyzing the global football player transfer market. In the transfer network, nodes are clubs linked by directed edges representing player transfers. The global football player transfer network is a small-world network with multiple loosely connected hubs. Clubs that act as hubs or brokers in the network usually achieve better domestic and international match performances. The results suggest that professional clubs should develop their scouting abilities and maintain exclusive player resources in order to achieve better match performances.

The ultimate goal of commercial organizations is to make profit. However, depending on various factors, football industry does not generate a comparable amount of revenue across the world. In some leagues, clubs could spend millions of euros on building a better team and profit from commercial activities. Yet, this strategy might not apply on clubs in leagues that attract less financial attention. Therefore, cultivating players with high potential and selling them to wealthier leagues is another viable way of generating profit for the clubs. No matter which strategy a club has to choose, in order to achieve success, the club must carefully select its position, particularly the coreness and brokerage properties, in the global player transfer network.

Meanwhile, the domestic and international movement of football players is merely a special case of the ongoing urbanization processes and global labor migration today. How the acquisition and loss of labors with different skill sets could affect the economic status of regions or nations is still an open question. We believe that the systematic perspective and network-based methods employed in this work can be further extended to study this question with a promising outcome.

References


Letter Reproduction using a Cellular Neural Network consisting of Simplified Neurons and Synapses fabricated by Thin-Film Transistors

Mutsumi Kimura†,‡, Ryohei Morita‡, Sumio Sugisaki‡, Tokiyoshi Matsuda‡, Tomoya Kameda†, and Yasuhiko Nakashima†

†Graduate School of Information Science, Nara Institute of Science and Technology
Takayama, Ikoma 630-0192, Japan
‡Department of Electronics and Informatics, Ryukoku University
Seta, Otsu 520-2194, Japan
Email: mutsu@rins.ryukoku.ac.jp

Abstract—Simplified neurons and synapses fabricated by thin-film transistors (TFTs) are proposed for cellular neural networks. A two-inverter two-switch circuit is used as a neuron, whereas only a transistor is used as a synapse. The neurons and synapses are fabricated by TFTs, which are promising for giant microelectronics. A cellular neural network consists of such neurons and synapses. Particularly in this presentation, it is observed that an alphabet letter is reproduced from a similar pattern by the cellular neural network. This function is available for letter recognition of hand-written letters.

1. Introduction

Cellular neural networks are neural networks where a neuron is connected to only neighboring neurons [1], hence exceedingly suited to integration of semiconductor circuits, and fundamental theory, operation principle, and potential applications, such as image processing [2] and pattern recognition [3], have been actively investigated using formal models and numerical simulation until now. However, actual hardware of cellular neural networks has been rarely reported [4], and we think this is because the conventional circuits of the neurons and synapses are rather complicated, even if the network architecture is extremely simplified. Therefore, an objective of this study is to simplify the neurons and synapse.

Thin-film technologies are promising for giant microelectronics having potential possibility for astronomical ultra-large-area integrated circuits [5]. Although TFTs have been widely utilized for flatpanel displays [6],[7], novel applications are ardently desired [8]. Therefore, the other objective of this study is to fabricate neurons and synapses by TFTs.

We are continuing to investigate artificial neural networks fabricated by TFTs [9]-[12]. In this study, simplified neurons and synapses fabricated by TFTs are proposed for cellular neural networks. A 2-inverter 2-switch circuit is used as a neuron, only a TFT is used as a synapse, and a cellular neural network consists of such neurons and synapses. Particularly in this presentation, it is observed that an alphabet letter “T” is reproduced from a similar pattern by the cellular neural network. This function is available for letter recognition of hand-written letters. It should be noted that the neurons, synapses, and cellular neural network were actually fabricated, and the experiment results were really obtained.

2. Neuron

Figure 1 shows the circuit diagram of the neuron. A 2-inverter 2-switch circuit is used as a neuron, where the two inverters and two switches are circularly connected. A inverter consists of a pair of n-type and p-type transistors, and a switch also consists of a pair of the transistors. The 2-inverter 2-switch circuit has the necessary functions we considered: generating a binary state and alternating the binary state by the input signal. The inverters generate a binary state, and the binary state is maintained when the switches are on, whereas the binary state is alternated when

Fig. 1. Circuit diagram of the neuron.
the switches are off and some input signal is received. The switches are periodically on and off in the entire cellular neural network. The two terminals are bi-directional, that is, work as both input and output terminals. One terminal is for positive logic, whereas the other terminal is for negative logic. This neuron circuit is fabricated by eight transistors, and the number of the synapse per connected neighboring neuron is two. The n-type transistors have the gate oxide thickness ($t_{ox}$) = 75 nm, gate width ($W$) = 100 $\mu$m, gate length ($L$) = 7.5 $\mu$m, lightly-doped drain length (LDD) = 0.75 $\mu$m, field-effect mobility ($\mu$) = 93 cm$^2$V$^{-1}$s$^{-1}$, and threshold voltage ($V_{th}$) = 3.6 V, whereas the p-type transistors have the same $t_{ox}$, same $W$, $L = 5 \mu m$, single drain structure, $\mu = 47 $ cm$^2$V$^{-1}$s$^{-1}$, and $V_{th} = -2.9$ V. The detailed explanations for the neuron are elsewhere [9],[10].

3. Synapse

Figure 2 shows the circuit diagram of the synapse. Only a TFT is used as a synapse, where the gate voltage is used as a control voltage to induce the characteristic shift or not, and the source and drain terminals are connected to the neurons. The control voltage is constantly applied in the entire cellular neural network. The transistor has the necessary functions we considered: sending the signal from a neuron to the neighboring neuron, merging the signals from the multiple neurons for the neuron, and controlling the synaptic connection strength on demand. The transistor sends the signal as an electric current. The conductance corresponds to the synaptic connection strength. The electric currents are easily added by bundling the variable resistors in parallel, which corresponds to merging the signals. The characteristic shift occurs owing to the characteristic degradation by the electric current, which
corresponds to controlling the synaptic connection strength. The n-type transistor has the same parameters as those in the neurons except for \( W = 5 \ \mu m \) and single drain structure to induce the characteristic shift.

4. Neural Network

Figure 3 shows the network architecture of the neural network. The cellular neural networks have 7×7 neurons including 3×3 input/output (I/O) neurons, to which the letter pattern is inputted and from which that is reversely outputted, and hidden neurons in between. A neuron is connected to the four neighboring neurons through two-type synapses, concordant and discordant synapses. The concordant synapse connects the same logics of the two neurons, that is, positive and positive logics or negative and negative logics, and tends to make the states of the two neurons the same, whereas the discordant synapse connects the different logics of the two neurons, that is, positive and negative logics, and tends to make the states of the two neurons different. Although the size of the cellular neural network is 5×5 mm², because most area are used for experimental evaluation, it can be reduced to a several thousandth.

5. Experiment method

In the learning stage, first, a control voltage of 15 V to induce the characteristic shift is uniformly applied to the TFTs for the synapses. A letter pattern of “T” is inputted to the terminals for the positive logic in the I/O neurons as an input pattern of the high (H) and low (L) voltages for several minutes. (Some cellular neural networks correctly worked after the letter pattern was inputted for one minutes, whereas some other ones worked after more than one minutes such as ten minutes.) A steady pattern of the neuron states is generated in the hidden neurons based on a normal theory of dynamics of neural networks. After a while, the synaptic connection strengths are changed. The conductance of the transistors of the concordant connections are kept the same when both neurons connected to the synapses are in the same states, and are impaired otherwise, whereas the conductance of the discordant connections are kept the same when both neurons are in the different states, and are impaired otherwise. This is because electric currents flow between the source and drain terminals owing to the voltage differences in the transistor for the concordant connections when both neurons are in the different states and discordant connections when both neurons are in the same states, and characteristic degradations are induced owing to the Joule-heating and hot-carrier degradations by the electric currents [13]-[15]. We call this a modified Hebbian learning rule [16]. In the recalling stage, finally, a control voltage of 10 V to avoid the characteristic shift is applied. A letter pattern with a slightly different part from “T” is initially inputted to the I/O neurons and immediately, namely less than 1 s, released, and an output pattern is automatically outputted from the I/O neurons. It is confirmed whether it is the same as the letter pattern of “T” inputted first.

6. Experiment result

Figure 4 shows the experiment result of the letter reproduction. Only the states in I/O neurons are shown. In the learning state, a letter pattern of “T” is inputted to the I/O neurons for several minutes. In the recalling stage, a letter pattern with a slightly different part from “T” is initially inputted to the I/O neurons and immediately released, and an output pattern is automatically outputted from the I/O neurons. The response time is less than 0.1 s. It is observed that an alphabet letter “T” is exactly reproduced by the cellular neural network.

Incidentally, some power are consumed especially in the synapses because direct current flows, which we will evaluate soon. Moreover, we are now trying other alphabet letters and would like to report in the near future. At that time, other architectures such as 7×11 neurons including 3×5 I/O neurons or 11×17 neurons including 5×8 I/O neurons may be suitable for English alphabet letters. Simultaneous recognition of two or more letters are the next step of this research. We are also now trying it and would like to report in the near future.

7. Conclusion

We are continuing to investigate artificial neural networks fabricated by TFTs. In this study, simplified neurons and synapses fabricated by TFTs were proposed for cellular neural networks. A 2-inverter 2-switch circuit was used as a neuron, whereas only a transistor was used as a synapse, and a cellular neural network consisted of such neurons and synapses. The neurons and synapses are fabricated by TFTs, which are promising for giant...
microelectronics. Particularly in this presentation, it was observed that an alphabet letter “T” is reproduced from a similar pattern by the cellular neural network. This function is available for letter recognition of hand-written letters.

References


Letter Reproduction Simulator for Hardware Design of Cellular Neural Network consisting of Neurons using Large-Scale Integration Chip and Synapses using on-Deposited Amorphous In-Ga-Zn-O Films

Tomoya Kameda†, Mutsumi Kimura†‡, and Yasuhiro Nakashima†

†Graduate School of Information Science, Nara Institute of Science and Technology
Takayama, Ikoma 630-0192, Japan
‡Department of Electronics and Informatics, Ryukoku University
Seta, Otsu 520-2194, Japan
Email: kameda.tomoya.kg0@is.naist.jp

Abstract—Recently, neural networks have been developed for variable purposes including image recognition and voice recognition. However, such neural networks based on software require many times of calculation and large quantity of consuming energy. Therefore, we are developing hardware of a cellular neural network (CNN) that features low power consumption, high-density integration of electron devices, and high functionality. In our hardware system, amorphous In-Ga-Zn-O (a-IGZO) films work as synapses because they have a potential to integrated astronomical number of devices and their characteristic is available for our leaning rule. Particular in this presentation, we assume to form the neuron circuits using a large-scale integration chip and synapse devices using a-IGZO films deposited directly on the neuron chip. We modeled a-IGZO film for CNN and implemented it into the simulator to determine the network architecture and device parameters. In this time, we succeeded in confirming that this CNN can learn two letters of Arabic numerals. Moreover, we estimated the time necessary for the learning.

1. Introduction

Neural networks are types of information systems that imitate biological neural circuits in living cells of human brains [1]. Therefore, they are promising for image processing, pattern recognition, etc., that human brains can do well. Neural networks can process many and complicated data by using unbelievable figures of neurons and synapses. For example, in a human brain, there are more than $10^{10}$ neurons and $10^{13}$ synapses that connects the neurons generally [2]. Neural networks also need astronomical number of neurons and corresponding number of synapses to calculate required operations fast and efficiently.

There exist several kinds of neural networks. Among the neural networks, cellular neural networks (CNN) are suitable for actual integration of electron devices, because each neuron is connected to only neighboring neurons and there are no connections between the neurons far away [3]. Further technologies on CNN preferable for higher-density integration have been also developed from the viewpoint of device hardware [4].

In order to design a CNN at the circuit level, we are evaluating the CNN by using a simulator. Particular in this presentation, we assume to form the neuron circuits using a large-scale integration (LSI) chip and synapse devices using a-IGZO films deposited directly on the neuron chip. We will introduce the CNN that we are studying, explain a letter reproduction simulator that we developed for the CNN, and show the simulation results.

2. Cellular Neural Network

2.1. Network Architecture

Figure 1 shows the network architecture of the CNN that we are studying and used for our simulator. In this architecture, we arrayed multiple neurons and connected each neuron to eight adjacent neurons through synapses. The neurons are in a state of either firing or non-firing according to the states of the adjacent neurons and the connection strength of the synapses placed between the

![Fig.1. Network Architecture of our Cellular Neural Network](image-url)
neurons. The synapses tend to make the states of the neurons connected through the synapses the same. When both neurons connected to the synapse are in firing states, the synaptic connection strength doesn’t change. On the contrary, when the neurons connected to the synapse are in different states, the synaptic connection strength decreases. Such deterioration of the synapses corresponds to the learning of the CNN. We call this method of the learning “Modified Hebb’s learning” [5].

2.2. Neuron Circuit

Figure 2 shows the neuron circuit. Figure 2 (a) shows the circuit architecture. We limited the necessary functions of the neuron to that a binary state is maintained by itself and altered by the input signals. We can make a neuron consist of only four transistors, therefore it is suitable for integration. Moreover, this simple architecture can be easily formed using an LSI chip. Figure 2 (b) shows the computer-aided design (CAD) layout, which we will develop in near future. Here, in addition to a circuit to connect adjacent neurons, I/O circuit is designed. The size of a neuron is only 100μm × 100μm. The light blue patterns are electrodes for the synapses. We can connect neurons through these electrodes.

2.3. Synapse Modeling

We assumed to use a-IGZO films as synapses. The conductance of the a-IGZO films changes when an electric current flows [6], if they are made using some fabrication process. We will deposit the a-IGZO films on all neurons as shown in Figure 3. The light blue patterns are electrodes, and the light yellow patterns are a-IGZO films. The parts of the a-IGZO films indicated by the deep yellow pattern work as synapses. The adjacent neurons are connected by the synapses. In consideration of the a-IGZO film size (2μm × 10μm) and thickness (70nm) and electrical characteristics, we can guess the resistance and deterioration speed of a-IGZO from the experimental results. The initial resistance is 3.75MΩ and the resistance increases about 1% in 4 seconds when the electric current flows.

3. Letter Reproduction Simulator

3.1. Neuron Arrangement

In this time, we use a CNN for letter reproduction. First, we set 25 × 25 neurons. Next, we put hidden neurons between all pairs of the adjacent I/O neurons. Figure 4 shows the arrangement of the I/O neurons and hidden neurons. Here, all neurons are indicated by 625 squares, the 12 × 12 matrix of the neurons surrounded by bold lines are I/O neurons, and other neurons are hidden neurons. As a result, we can input letter images converted to the binary format of 12 × 12 pixels. In the learning stage, a pixel pattern corresponding to an input letter is inputted to the I/O neurons as firing states. For example, when a letter of “0” is inputted, firing states are inputted to the I/O neurons indicated by the red square, whereas non-firing states are inputted to the other I/O neurons.
3.2. Simulation Algorithm

The letter reproduction simulator simulates the CNN based on the network architecture and synapse model mentioned above, which has also already reported in a previous conference [7]. Figure 5 shows the simulation algorithm of the letter reproduction simulator. First, a pixel pattern corresponding to an input letter is inputted to I/O neurons as firing states. Afterwards, the states in the hidden neurons are calculated using the majority rule of the neighboring neurons with consideration of the synaptic connection strengths corresponding to the conductance of the a-IGZO films. Next, after the all the states in the neurons are settled, the modified Hebb’s learning are done, namely, when the neurons connected to the synapse are in different states, the synaptic connection strength decreases based on the deterioration model of the a-IGZO films. After that, a pixel pattern slightly distorted from the input pattern is inputted to I/O neurons as firing states and immediately removed. Afterwards, the states in all neurons including the I/O and hidden neurons are calculated. Next, it is checked whether the output pattern from the I/O neurons becomes the original inputted pattern. Finally, these procedures are repeated for multiple input patterns.

4. Simulation Result

4.1. Letter Reproduction

Figure 6 shows the input pattern for the learning and that for the reproduction slightly distorted from the former. Figure 7 shows the simulation results of the letter reproduction. Figure 7 (a) show the states of neurons and synapses after learning, whereas figure 7 (b) show the state of neurons and synapses after reproduction, when this CNN was able to reproduce two letters. It was confirmed that this CNN can reproduce two letters. The color deepness of the red small squares corresponds to the synaptic connection strength. It was confirmed that the firing patterns after the reproduction are completely the same as the input pattern for the learning. These results mean that this CNN can reproduce two letters.
4.2. a-IGZO film Deterioration

In this simulation, we recorded the resistance of the synapses that increases most, which pointed by the arrow in Fig. 7. Figure 8 shows the resistance of the synapses that increases most. When this resistance increased from 3.75MΩ to 11.25MΩ, this CNN was able to reproduce two letters. This result indicates that it takes about 446 seconds, when we use this a-IGZO film. Although we succeeded in confirming that this CNN can learn two letters as mentioned above, it takes long time and it should be shortened by finding novel structures or fabrication process of a-IGZO films to enhance the deterioration of the conductance.

5. Conclusion

We are developing the CNN system using an LSI chip as neuron circuits and a-IGZO films as synapses, and we developed the letter reproduction simulator for the CNN. We modeled the deterioration phenomenon of the a-IGZO films and implemented it into the simulator. It is confirmed using the simulator that this CNN can reproduce two letters of Arabic numerals. Moreover, we evaluated the time necessary for the learning two characters. We believe that these results indicate the future possibility of the CNN using a-IGZO films as the synapses to realize astronomical large-scale brain-like integrated systems, and we will continue to develop them, which will be reported in the near future.

In this CNN, the neuron, synapse, and network architecture are quite different from those in the conventional one. Therefore, we might have to make some unique representations. Moreover, we would like to formalize the behaviors of these processing elements and present the dynamics and local interaction rules in the near future.

References


Multilayer Perceptron Including Different Amplitude Random Noise

Chihiro Ikuta† Yoko Uwate‡ and Yoshifumi Nishio§

†Course of Electrical Engineering, Dept. of Creative Technology Engineering,
National Institute of Technology, Anan College
265 Aoki, Minobayashi, Anan, Tokushima 774-0017, Japan
‡Dept. Electrical and Electronics Engineering, Tokushima University
2–1 Minami-Josanjima, Tokushima 770-8506, Japan
Email: ikuta@anan-nct.ac.jp, uwate, nishio@ee.tokushima-u.ac.jp

Abstract—In this study, we introduce different amplitude random noise to a multilayer perceptron (MLP). We divide the neurons in hidden layer to some groups according to value of neuron output. Each neuron group is input the different amplitude random noise. When a neuron group has a large output, this neuron receives a large amplitude noise. When a neuron group has a small output, this neuron receives a small amplitude noise. The group member is dynamically changed during the learning because the neuron output changes with the learning. By simulations, we confirm that the proposed MLP performance is better than the standard MLP and MLP having simple noise input method. Moreover, we show the parameter dependency of the proposed MLP by changing the number of groups and the noise amplitude.

1. Introduction

Multilayer perceptron (MLP) is one of feed forward neural network which is applied to pattern recognition, data mining, and so on. In the MLP, neurons make some layers and connect with neurons in other layer. The MLP output is decided by the weight of connection between the neurons. In general, we use the back propagation algorithm (BP) [1] for a learning of weight of connection. The BP uses the steepest decent method, thus the network is often traps into local minimum. Then, the network cannot escape out from the local minimum. For a problem of local minimum, some methods are proposed such as a pre-training, a noise input, and an annealing [2][3].

In this study, we propose an MLP including different amplitude random noise for escaping out from the local minimum. We use the MLP of three layers and input a uniformed random noise to the inner state of the neurons in hidden layer. In general, the uniformed random noise is not efficiency to the escaping out from the local minimum. If amplitude of the random noise is larger, the network cannot learn the supervised signal, because many neuron outputs become random. On the other hand, if amplitude of the random noise is smaller, the network falls into local minimum. In our method, we divide some groups according to the descending order of output value. The neuron having large output value obtain the large amplitude random noise, and the neuron having small output value obtains small amplitude random noise. By this method, the network can learns the supervised signal under influence of the random noise and escaping out from the local minimum. By the simulation, we confirm that the different amplitude random noise is efficiency to the MLP learning, and dependency of the dividing number of neuron groups.

2. Proposed Method

The proposed MLP is shown in Fig. 1. We divide the neurons in hidden layer to some groups according to the output value of neuron. Each group has different amplitude random noise, and this noise is input into the neurons. In the same group, the neuron has same amplitude random noise. The group of the neuron having a large output obtains the large amplitude noise. On the other hand, the group of the neuron having a small output obtains the small amplitude noise.

Figure 1: Proposed MLP.

2.1. Neuron Updating Rule

A standard neuron updating rule is described by Eq. (1).

\[ y_i(t+1) = f \left( \sum_{j=1}^{n} w_{ij}(t)x_j(t) - \theta_i(t) \right), \]  

(1)
where $y$ is an output of the neuron, $w$ is a weight of the connection, $x$ is an input of the neuron, and $\theta$ is an excitation threshold of the neuron. In this equation, the weight of the connection and the threshold of the neuron are learned based on the BP algorithm. Next, we show the updating rule of the neuron with the uniformed random noise in Eq. 2. This updating rule is used for the neuron in the hidden layer.

$$y_i(t+1) = f \left( \sum_{j=1}^{n} w_{ij}(t)x_j(t) - \theta(t) + \alpha_n R(t) \right), \quad (2)$$

where $\alpha$ is amplitude of the uniformed random noise, and $R$ is a uniformed random noise. Here, the uniformed random noise is given by Mersenne Twister (MT) which was proposed by Matsumoto and Nishimura [4]. The uniformed random noise gives the energy to the network and helps escaping out from the local minimum.

2.2. Different Amplitude Random Noise

The amplitude of the input random noise ($\alpha$) decreases at an exponential rate. The $\alpha$ is described by Eq. (3).

$$\alpha_n = 0.8^{n-1} \quad (n = 1, 2, \ldots, N). \quad (3)$$

Figure 2 is an overview of the proposed noise input method. We divide the neurons to some groups according to the value of neuron output. In this figure, six neurons are divided to three groups. L means that neuron has large output value. M means that the neuron has middle output value. S means that the neuron has small output value. Each group has different amplitude random noise. In the same layer, the same amplitude random noise. During the learning, the group is dynamically changed, because the value of neuron output changes with the learning.

3. Simulation

In this section, we show the simulation results. We use four kinds of MLPs which are;

1. Standard MLP.
2. MLP with uniformed random noise.
3. MLP including different amplitude random noise (the neuron group is fixed).
4. MLP including different amplitude random noise (the neuron group is changed according to the value of neuron output).

The standard MLP (1) does not have the noise, thus this MLP reduces error faster, however it often falls into local minimum. The MLP with uniformed random noise (2) has the uniformed random noise in the inner state of neurons in hidden layer. In the MLP including different amplitude random noise (3), we fix the neuron groups during the simulation, thus the neuron receives constant noise amplitude. The MLP including different amplitude random noise (4) is proposed MLP. The number of neurons in input layer, hidden layer and output layer is 2, 40 and 1, respectively. The number of iterations is 50000. We obtain the result from 100 trials from different initial weight of connection. We use the mean square error (MSE) to the error function which is described by Eq. (4).

$$MSE = \frac{1}{N} \sum_{n=1}^{N} (T_n - O_n)^2, \quad (4)$$

where $N$ is the number of learning datum, $T$ is a target value, and $O$ is an output of MLP. We use the two spiral problem (TSP) for the task of the MLPs. The TSP is famous task for the ANN and has high nonlinearity, thus the standard MLP often falls into local minimum [5][6]. We give the coordinates of each point to the neuron in input layer, and MLP learns the classification of the spiral point. In this simulation, the two spirals are constructed from 130 points.

3.1. Comparison learning performance of four MLPs

Firstly, we show the learning performance of each MLP. We compare the four measures which are average, minimum, maximum, and standard deviation (Std. Dev.). The
learning performance is shown in Table 1. From this table, every MLP reached to 0.00 in minimum error, thus the MLP can solve this task when optimal initial weight is set. However, the MLP often falls into local minimum. Especially, the average error and the maximum error in the standard MLP (1) is the largest of all. Other MLPs reduce the error from the standard MLP. The MLP with uniformed random noise (2) is similar to the MLP with different amplitude random noise (3). From this result, the different amplitude is not efficiency to the learning performance when we simply introduce the different amplitude random noise to the MLP. The proposed MLP (4) reduces the error than the MLP with uniformed random noise (2) and MLP including different amplitude random noise (3). We divide the neurons in hidden layer to some groups according to the value of neuron output, thus the unadapted amplitude noise is not input to the neuron.

3.2. Dependency of the number of groups and amplitude of the random noise

Next, we change the number of groups in divided neurons in hidden layer and the maximum amplitude of the random noise. Firstly, we show the result of the MLP with uniformed random noise. In the MLP with uniformed random noise, we do not divide the groups, thereby we only show the result of changing maximum amplitude of random noise. In this case, the MSE decreases to the amplitude 0.01, after that the MSE increases with the maximum amplitude. From this result, the allowable noise amplitude is small in the MLP learning. If the noise amplitude increases, the noise disturbs the learning.

Secondly, we show the result of the MLP including different amplitude random noise (3) in Fig. 5. We show three graphs in this figure. This number means the number of groups of the neurons in hidden layer. In the MLP including different amplitude random noise (3), the change of MSE becomes small. The MSE becomes small when the noise amplitude is between 0.1 and 0.4. Moreover, we cannot see the high dependency of the number of groups. Each neuron have same amplitude from start of learning to end of learning, thereby this influence is almost same as the MLP with uniformed random noise (2).

Finally, we show the result of the proposed MLP (4) in Fig. 6. In this case, the dependency of the number of groups of neurons is large. The learning performance is high when we divide five groups. And also, it has high learning performance in wide range of the maximum amplitude of random noise.
The curve which we divide 2 groups is similar to the result of the MLP including the different amplitude random noise (3). We consider that this group includes many neuron which has various values of output, thereby the influence of divide group becomes small.

![Figure 6: Parameter dependency of proposed MLP.](image)

4. Conclusion

In this study, we have proposed the MLP including different amplitude noise. We input the uniformed random noise to the inner state of the neuron in hidden layer. Then, the neurons in hidden layer are divided to some groups according to the output value of neurons. The neurons in the same layer receives same amplitude noise. The amplitude of input noise becomes small when the neuron output becomes smaller. By the simulations, we confirmed that the proposed MLP has better learning performance than other MLPs. Moreover, we show the parameter dependency of the MLPs. From this result, the learning performance improves by dividing the groups of neurons in hidden layer according to the value of neuron output.

References


Modified Inductive Search for Solving Global Optimization Problems

Hideo Kanemitsu†

†Hokkaido University of Education;
I–2 Hachiman-cho, Hakodate, Hokkaido 040-8567, Japan.
Email: †kanemitsu.hideo@h.hokkyodai.ac.jp

Abstract—
Inductive Search for solving global optimization problems has attracted much attention because it showed the best performance at the 1st ICEO. However, since details of this method are not clear, the method has not received much attention. We investigated details of the method and implemented the method. We propose a modified inductive search by using a deterministic one-dimensional global search. Finally, we evaluate the performance of the implemented method and that of proposed method.

1. INTRODUCTION

Global optimization problem: “(global) minimize \( f(x) \equiv f(x_1, \ldots, x_n) : \mathbb{R}^n \to \mathbb{R} \) under the constraint: \( x \in S \subset \mathbb{R}^n \)” is widely formulated as mathematical models and is applied in many fields. Many methods for solving continuous global optimization problems have been proposed[1], and these methods are classified into deterministic framework, stochastic framework and heuristic framework.

The deterministic framework[2] repeatedly divides a given region into subregions and selects a subregion in which a global optimum is included. The stochastic framework[3] involves random sampling or a combination of random sampling and local search[4]. On the other hand, the heuristic framework (e.g., SA[5], GA[6], PSO[7]) have been intensively since the latter half of the 1980s and has been applied many fields. However, most of these methods have no guarantees to find a global optimal solution.

Searching spaces of all of those frameworks exponentially increase with increase in the number of dimensions in the problem (P). This phenomenon, known as the “curse of dimensionality”, led to the abandonment of those search methods in favor of ones using some a priori knowledge or priori structure of the function f.

Inductive search was proposed[8] at the 1st international contest on evolutionary optimisation and the search has achieved an best result in this contest[9]. However, since details of the method were not clear, the method has not attracted much attention recently.

The purpose of this paper is to introduce the original inductive search and to reconstruct the modified inductive search using our univariate global search[13].

The remainder of the paper is organized as follows. A formulation and assumptions of the problem are given in Sect. 2. In Sect. 3, the algorithm of the inductive search is introduced. An algorithm of a deterministic inductive search using a one-dimensional global search is shown in Sect. 4. Finally, concluding remarks are given in Sect. 5.

2. PRELIMINARIES

A global optimization problem that minimizes (min.) an objective function \( f : \mathbb{R}^n \to \mathbb{R} \) subject to (s.t.) constrains \( D^p \subset \mathbb{R}^n \) is formulated as follows

\[
\text{(P)} \begin{cases}
\min. : & f(x) \equiv f(x_1, x_2, \ldots, x_n), \\
\text{s.t. :} & (x_1, x_2, \ldots, x_n) \in D^p \equiv \prod_{i=1}^n [L_i, U_i] \subset \mathbb{R}^n.
\end{cases}
\]

In this problem we assume that objective function f is a Morse function, that is, f is 2nd continuous differentiable and Hessian matrix \( \nabla^2 f(x^*) \) at critical point \( x^* \) (s.t. \( \nabla f(x^*) = 0 \)) is non-degenerate (i.e., \(|\nabla^2 f(x^*)| \neq 0 \)). Then, the following two properties hold: a) all local minima of the problem (P) are isolated and b) f has a convex region around a critical point \( x^* \).

3. INDUCTIVE SEARCH

3.1. Algorithm and its Implementation

The most novel idea of the method is to solve a subproblem \((P^k)\) \((k = 1, 2, \ldots, n)\) inductively by increasing the number of variables. The pseudo-C++ implementation of the overall structure of the idea is given as follows.

```
1 #include "t1.c" // t1: Sphere nv. problem
2
3 int iter_count=0, int n;
4 for (int i=0; i<n; i++) oracle(i+1);
5 }
```

Fig.1: Main code for calling inductive search oracle(..).

In the above code, the argument \( i+1 \) of oracle\((i+1)\) denotes the number of dimensions on procedure oracle, and \( i \) increases by \( 1 \) until \( n - 1 \). For a problem, this is easy to achieve even by till treating them as black boxes, because the test function is defined in terms of two parameters, the number of dimensions and a vector of the input variables: \( f(x_1), f(x_1, x_2), \ldots \)

1Later codes are simplified codes of the original code[1, 2] with maintenance of logical structure.
At a later stage, i.e., when solving $f(x_1, \ldots, x_i)$, the oracle can “update” a previous answer by changing the values of $x_i$ to $x_{i-1}$. This is necessary because when the oracle solves $f(x_1, \ldots, x_{i-1})$, it has no knowledge of how this function will be updated to $f(x_1, \ldots, x_i)$.

The 1st line of Fig.1 is an include file for the Sphere problem, and the C-code of the file "t1.c" is shown in Fig.2.

```c
#define STP_C 0
#define LEARN 0
#define L -5.0
#define U 5.0
float f(float *x, int nv) {
  int i; float S;
  for (S=0.0, i=0; i<nv; i++) S+=x[i]*x[i];
  return (S);
}
```

**Fig.2.** Content of C-code: t1.c for Sphere problem

Bilchev [8] proposed a simple (deterministic) version of oracle that obtained very good results in the 1st ICEO test problems. The pseudo-C++ implementation of the basic algorithm is shown in Fig.3.

```c
void oracle(int nv) {
  int count=0;
  float xmin,FMIN=1e30,xmin,XMIN;
  void oracle(int nv) {
    float bren(float ax, float bx, float cx, float (*f1v)(float), float TOL, *xmin);
    ax=I.L; cx=I.U; bx=(ax+cx)/2;
    S=Pop.max(); Pop.del_item(S); I=Pop.inf(S);
    while (1) {
      Pop.insert(cx-ax,I);
    } /* Global learning */
    while (1) {
      S=Popp.max(); Pop.del_item(S); I=Pop.inf(S);
      ax=I.L; cx=I.U; bx=(ax+cx)/2;
      fmin=bren(ax,bx,cx,f1v,TOL,&xmin);
      if (fmin < FMIN) { FMIN=fmin; XMIN=xmin; }
      /* Global learning */
      while (1) {
        S=Popp.max(); Pop.del_item(S); I=Pop.inf(S);
        ax=I.L; cx=I.U; bx=(ax+cx)/2;
        fmin=bren(ax,bx,cx,f1v,TOL,&xmin);
        if (fmin < FMIN) { FMIN=fmin; XMIN=xmin; }
        /* Global learning */
        count++;
      }
      if (count > STP_C) break;
    }
    /* Local learning */
    if ( (LEARN) && (nv > 1) ) local_learn(fl,x,nv);
    xmin=XMIN;
  }
}
```

**Fig.3.** An implementation of the oracle.

In this code, the 5th, 7th and 10th lines denote services of LEDA [12]. If STP_C = 0, then the 11th line is equivalent to “ax=I.L; cx=I.U; bx=(ax+cx)/2;” from the 4th and 6th lines. In this case, these LEDA steps are unnecessary.

**S1) Global learning,** which is a search for a better solution at the current dimension than the previous best solution(s);

**S2) Local learning,** which an $nv$-(sub) dimensional local search: losrch( ).

In this implementation, the global learning is a series of calls Brent’s local optimizer routine: bren( ), and its external specification of the routine is as follows: bren( ) consists of two main steps:

1. float flv(float y) {
   x[nvdim-1] = y;
   float res = f(x,nvdim);
   return res;
}

**Fig.4.** An auxiliary routine of a one-dimensional function

### 3.2. Problems of an Inductive Search

Since the inductive search has not been followed, a pseudo-code is difficult to implement for the following reasons.

1. These codes use many global constants (e.g., STP_C, LEARN, L and U) and variables (e.g., nv and $x_1$ in Fig.4).
2. Searching regions L and U are all fixed for any test problem of files “tj.c” ($j = 1, 2, \ldots, 5$).
3. Oracle( ) needs steps supported by LEDA [12].

If STP_C = 0, then the above LEDA-steps are unnecessary and global learning is not performed.

### 4. MODIFIED INDUCTIVE SEARCH

#### 4.1. One-Dimensional Global Search Algorithm

The algorithm finds the minimum $x^*_i$ of $i$-th variables and its function value for an objective function $f_1(x)$ at the $i$-th variable on a closed interval $D_i = [L_i, U_i]$ for a given step size $h$ and an already found global minimum $x^*_{i-1} = (x^*_{i-1}, \ldots, x^*_{1})$. The steps of the algorithm are as follows.

$$ (f^*_i, x^*_i) \leftarrow Go_{1DimSrch}(f_i, D_i, h, x^*_{i-1}) $$

**GL1. [Initialize]**

$$ x_0 \leftarrow 0; \quad F_0 \leftarrow 0; \quad N \leftarrow [(U_i-L_i)/h]; \quad f^{00} \leftarrow f_1(L_i); \quad f^{01} \leftarrow f_1(L_i+h, x^*_{i-1}); \quad f_i^* \leftarrow \infty; $$

**GL2. [Find three neighboring points with bracketing of a local minimum]**

for $j = 2$ to $N$ do

$$ x^{j0} \leftarrow L_i + j \cdot h; \quad f^{j0} = f_1(x^{j0}, x^*_{i-1}); $$

if $f^{j0} \geq f^{j1}$ and $f^{j0} \geq f^{j2}$ then

$$ X_0 \leftarrow F_0 \cup \{ (x^{j0}, x^{j1}, x^{j2}) \}; $$

$$ F_0 \leftarrow F_j \cup \{ (f^{j0}, f^{j1}, f^{j2}) \}; $$

if $f^{j0} < f^{j2}$ then $f_i^* \leftarrow f^{j1}; \quad x^*_i \leftarrow x^{j0}; \quad \text{fi.}$

**GL3. [Apply local minimization]**

Apply univariate local minimization: LoMin1v( ).

$$ (f^*_i, x^*_i) \leftarrow LoMin1v(F_i, x_0, f_i^*, x^*, TOL); $$

**Fig.4.** An auxiliary routine of a one-dimensional function

In a one-dimensional global search Go_1DimSrch( ), the following property holds [13].

**Property 1** Let the lower unimodal region of the global minimum $x^*_i$ of $f(x)$ on an interval $[L_i, U_i]$ be $R_1(x^*_i)$, where $R_1(x^*_i)$ is defined as the maximum region. Then, if $h \leq 1/2 \cdot \min(x^*_i - a, b - x^*_i)$ holds, the algorithm Go_1DimSrch( ) always finds the global minimum $x^*_i$ of function $f(x)$.

An example of unimodal region $R_1(x^*_i)$ and step size $h$ is shown in Fig.5.
4.2. Modified Inductive Search Algorithm

The algorithm finds the global minimum \( x^{**} \) and its function value \( f^{**} \) for an objective function \( f(x) \) of \( n \)-variables with searching region \( D^n \equiv \prod_{j=1}^n D_j \equiv [L_j, U_j] \) for a given step size \( h \). The steps of the algorithm are as follows.

![Fig. 5: Unimodal region Ru(\( x^{**} \)) and step size h](image)

\[
\min \{ f(x) | x \in [L, U] \}, \quad s.t.: \quad x \in \mathbb{R}^n
\]

G1. [Perform one-dimensional global search] for \( n = 1 \) to \( n \) do

\[
(f^{**}, x^{**}) \leftarrow \text{Go}_n(\text{DimSrch}(f, n, D^n, h, \text{TOL}));
\]

G2. [Apply local minimization]

Apply local minimizer: \( \text{LoMin}(\cdot) \) to starting point \( x^{**} \) and its function value \( f^{**} \);

\[
(f^{**}, x^{**}) \leftarrow \text{LoMin}(f, n, f^{**}, x^{**}, \text{TOL});
\]

return \((f^{**}, x^{**})\);

4.3. Comparisons between the Original Inductive Search and Our Modified Search

We show comparisons between the original inductive search algorithm (O) and our algorithm (M) as follows.

\[
\text{O: Number of calling oracle(\( \cdot \)) (see Fig.1): } n
\]

\[
\text{O: Number of calling bren(\( \cdot \)) (see Fig.3): } n \times (\text{STP.C} + 1)
\]

\[
\text{M: Number of calling Go}_n(\text{DimSrch}(\( \cdot \), \( \cdot \), \( \cdot \), \( \cdot \), \( \cdot \))); \quad n
\]

\[
\text{O: Number of calling localLearn(\( \cdot \)), \( \text{LEARN} \): } n \times \text{LEARN,}
\]

where the value of \( \text{LEARN} \) is 0 or 1.

(M) Number of calling \( \text{LoMin}(\cdot, \cdot) \) : 1

5. Benchmark Functions of 1st ICEO[9]

The following five objective functions with two cases of dimensions \((n = 5, 10)\) at the 1st ICEO are presented.

1. Sphere (unimodal, separable):

\[
\min \{ f(x) | x \in [-5, 5]^n \}, \quad s.t.: \quad x \in \mathbb{R}^n
\]

2. Griewank (multimodal with convex skeleton, non-separable):

\[
\min \{ f(x) = \frac{1}{4000} \sum_{i=1}^{n} (x_i - s_i)^2 - \prod_{i=1}^{n} \cos \left( \frac{x_i - s_i}{\sqrt{n}} \right) + 1, \}
\]

\[
\text{s.t.: } \quad x \in [-600, 600]^n; \quad s = 100, \quad n = 5, 10.
\]

3. Shekel (multimodal, non-separable):

\[
\min \{ f(x) = - \sum_{i=1}^{n} \frac{1}{|x-a_i|^2 + c_i}, \}
\]

\[
\text{s.t.: } \quad x \in [0, 10]^n; \quad n = 5, 10.
\]

4. Michalewicz (multimodal, separable):

\[
\min \{ f(x) = - \sum_{i=1}^{n} \sin(x_i \sin(\sqrt{2n})) \},
\]

\[
\text{s.t.: } \quad x \in [0, \pi]^n; \quad n = 5, 10.
\]

5. Langerman (multimodal, non-separable):

\[
\min \{ f(x) = - \sum_{i=1}^{n} c_i \left( e^{0.1|x-a_i|^2} - \cos(\pi |x-a_i|) \right), \}
\]

\[
\text{s.t.: } \quad x \in [0, 10]^n; \quad n = 5, 10.
\]

5. Implementation

The inductive search and its modified search are implemented in programming language C (MinGW g++ 3.4.5). All numerical experiments for which results are shown in this paper were carried on an Lenovo note PC Think Pad X250 (2.6GHz Intel Core i7-5600) with double precision.

Setting parameters of inductive searches for five test functions are shown in Table 1.

<table>
<thead>
<tr>
<th>Function</th>
<th>STP.C</th>
<th>LEARN</th>
<th>h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Griewank</td>
<td>0</td>
<td>0</td>
<td>1.2</td>
</tr>
<tr>
<td>Shekel</td>
<td>1</td>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>10</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>Langerman</td>
<td>2</td>
<td>1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

In this table, if the parameter STP.C = 0, then global learning becomes a local search because of one-dimensional local learn(\( \cdot \)) executing only one for each calling oracle(\( \cdot \)), (\( n=1,2,...,n \)) from sect.3.2. From sect.4.3, the number of calling bren(\( \cdot \)) is \( n \) in this case.

5.5. Comparison between Inductive Searches and the Other Methods

Comparisons between the number of callings in the original inductive searches and the number of callings in the other methods is shown in Table 2.

<table>
<thead>
<tr>
<th>Function</th>
<th>( n )</th>
<th>( N^5/N^0 )</th>
<th>( N^5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>5</td>
<td>25/20</td>
<td>243-12,218</td>
</tr>
<tr>
<td>Sphere</td>
<td>10</td>
<td>50/40</td>
<td>243-85,692</td>
</tr>
<tr>
<td>Griewank-1</td>
<td>5</td>
<td>68/41</td>
<td>5,765-2,977,996</td>
</tr>
<tr>
<td>Griewank-2</td>
<td>10</td>
<td>140/79</td>
<td>6,446-2,110,889</td>
</tr>
<tr>
<td>Shekel</td>
<td>5</td>
<td>415/74</td>
<td>6,318-451,992</td>
</tr>
<tr>
<td>Shekel</td>
<td>10</td>
<td>853/120</td>
<td>6,075-4,440,948</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>5</td>
<td>183/120</td>
<td>1,877-60,219</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>10</td>
<td>448/501</td>
<td>10,083-20,333,341</td>
</tr>
<tr>
<td>Langerman</td>
<td>5</td>
<td>471/176</td>
<td>4,131-232,496</td>
</tr>
<tr>
<td>Langerman</td>
<td>10</td>
<td>892/372</td>
<td>26,973-15,727,653</td>
</tr>
</tbody>
</table>
5.4 Result for Original Inductive search in ICEO and Our Implemented Inductive Search

Benchmark result of the original inductive search in the 1st ICEO[9] and our inductive search are shown in Table 3.

Table 3: Benchmark results of two inductive searches w.r.t. obtained minimal function values and the number of callings (\( f^{*}(N) \); original and \( f^{**}(N) \); our implementation).

<table>
<thead>
<tr>
<th>Function</th>
<th>( n )</th>
<th>( f^{*}(N) )</th>
<th>( f^{**}(N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>5</td>
<td>3.88 \times 10^{-17} (20)</td>
<td>0.25 (25)</td>
</tr>
<tr>
<td>Sphere</td>
<td>10</td>
<td>7.10 \times 10^{-15} (40)</td>
<td>0.50 (50)</td>
</tr>
<tr>
<td>Griewank</td>
<td>5</td>
<td>7.99 \times 10^{-8} (41)</td>
<td>1.09 \times 10^{-3} (68)</td>
</tr>
<tr>
<td>Griewank</td>
<td>10</td>
<td>1.31 \times 10^{-6} (79)</td>
<td>6.20 \times 10^{-1} (140)</td>
</tr>
<tr>
<td>Shekel</td>
<td>5</td>
<td>-10.327 (74)</td>
<td>-10.400 (415)</td>
</tr>
<tr>
<td>Shekel</td>
<td>10</td>
<td>-10.101 (120)</td>
<td>-10.208 (853)</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>5</td>
<td>-4.69 (120)</td>
<td>-4.49 (183)</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>10</td>
<td>-9.66 (501)</td>
<td>-8.75 (448)</td>
</tr>
<tr>
<td>Langerman</td>
<td>5</td>
<td>-1.499 (176)</td>
<td>-0.965 (471)</td>
</tr>
<tr>
<td>Langerman</td>
<td>10</td>
<td>-1.499 (372)</td>
<td>-0.076 (892)</td>
</tr>
</tbody>
</table>

The original inductive search finds global minima for all ten problems, but our implemented inductive search cannot find global minima for five problems.

Since Griewank’s problem is a multimodal function, and the parameter \( \text{STEP} = 0 \) and \( \text{LEARN} = 0 \) (i.e., global and local learning of an inductive search is not a valid.), our implemented search fails to find the global minimum \( x^{*} = (1, \ldots, 1) \). On the other hand, Original code’s inductive search function set \( s = 0 \), the original search function \( \text{c}_N \) can find the global minimum \( x_{\text{min}} = 0 \) because \( bx = (L+U)/2 = 0 \).

5.5 Results of Inductive Search and Modified Inductive Search

The result between our implemented inductive search and our modified inductive search are shown Table 4.

Table 4: Benchmark results of implemented inductive search and modified inductive searches w.r.t. obtained minimal function values and the number of calling (\( f^{**}(N) \); inductive search and \( f^{\prime\prime}_{\text{M}}(N) \); modified inductive search).

<table>
<thead>
<tr>
<th>Function</th>
<th>( n )</th>
<th>( f^{**}(N) )</th>
<th>( f^{\prime\prime}_{\text{M}}(N) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>5</td>
<td>0 (25)</td>
<td>0 (61)</td>
</tr>
<tr>
<td>Sphere</td>
<td>10</td>
<td>0 (50)</td>
<td>0 (121)</td>
</tr>
<tr>
<td>Griewank</td>
<td>5</td>
<td>1.09 \times 10^{-1} (68)</td>
<td>3.20 \times 10^{-1} (5,043)</td>
</tr>
<tr>
<td>Griewank</td>
<td>10</td>
<td>6.20 \times 10^{-1} (140)</td>
<td>2.12 \times 10^{-1} (10,167)</td>
</tr>
<tr>
<td>Shekel</td>
<td>5</td>
<td>-10.404 (415)</td>
<td>-10.404 (315)</td>
</tr>
<tr>
<td>Shekel</td>
<td>10</td>
<td>-10.208 (853)</td>
<td>-10.208 (810)</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>5</td>
<td>-4.490 (183)</td>
<td>-4.689 (169)</td>
</tr>
<tr>
<td>Michalewicz</td>
<td>10</td>
<td>-8.745 (448)</td>
<td>-9.160 (407)</td>
</tr>
<tr>
<td>Langerman</td>
<td>5</td>
<td>-0.01 (362)</td>
<td>-0.820 (355)</td>
</tr>
<tr>
<td>Langerman</td>
<td>10</td>
<td>-3.4 \times 10^{-3} (807)</td>
<td>-0.813 (682)</td>
</tr>
</tbody>
</table>

The number of function evaluations of an implemented inductive search for Sphere and Griewank functions is very small, because the parameters of Sphere and Griewank functions are set to \( \text{STEP} = 0 \) and \( \text{LEARN} = 0 \) (i.e., global and local learning of an inductive search is not a valid.).

Our modified inductive search finds better minima for six problems and with a smaller number of function evaluations for six problems than does the implemented inductive search.

6. Conclusion

In this paper, we showed a clearer and more detailed algorithm based on C-code or C++-code of an inductive search by Bilchev’s implementation. We propose an algorithm for a modified inductive search using a one-dimensional global search and we evaluated the algorithm. Both algorithms become deterministic methods because a random number generator is not used. Moreover, the one-dimensional global search of our modified method has a theoretical guarantee that the method can find a global minimum. Numerical examples show that our modified method can be reliably find a global minimum.

References

A Chaotic Search with the Effect of Wide Range Search for Solving QAP

Hikaru Ohnishi†, Yutaka Shimada‡, Kantaro Fujiwara‡ and Tohru Ikeuchi†‡

†Department of Management Science, Graduate School of Engineering
‡Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6–3–1 Niijuku, Katsushika-ku, Tokyo 125–8585, Japan

Abstract—One of the most important issues in engineering and science is to develop algorithms for finding good approximate solutions of NP-hard combinatorial optimization problems. In this paper, we proposed a new algorithm for solving Quadratic Assignment Problem (QAP) by chaotic neural dynamics. The proposed algorithm introduced a modified assignment of the neuron, which means that while the conventional method assigns neurons to the pair of facility and location, the proposed method assigns neurons to the location. In addition, we changed the effect of an external input: even if an exchange is bad, we applied a strong input to the chaotic neural dynamics. This effect enables a wide range search. As a result, our algorithm can find good solutions even though the number of neurons are reduced.

1. Introduction

In our daily life, many optimization problems exist, for example, scheduling, vehicle routing, facility location problem and so on. It is important to obtain possibly optimal solutions of these problems, because the cost can be reduced. However, it is very hard to obtain optimal solutions of such problems, because these problems are often classified into nondeterministic polynomial time solvable (NP)-hard problems. Therefore, we need to develop approximate algorithms to obtain near optimal solutions of these problems in a reasonable time frame.

On the other hand, several approximate algorithms are proposed for solving these NP-hard problems. One of the well known methods to solve these problems is a heuristic method, for example the 2-exchange method of Quadratic Assignment Problem (QAP). However, the heuristic methods such as the 2-exchange method are generally trapped into local minima. Due to this reason, many methods to escape from the local minima have also been proposed: for example, a tabu search[2, 3], a genetic algorithm[4], chaotic neural dynamics[5, 6] and so on. In this paper, we improve the method of using chaotic neural dynamics that we have already proposed in Ref.[7]. The method that we have proposed controls the 2-exchange method by using the chaotic neural dynamics. Compared with the method in Ref.[6], our proposed method reduced the number of neurons. In addition, to improve the performance of the method, we introduced a new strategy. Namely, we changed the effect of an external input: even if an exchange is bad, we applied a strong input to chaotic neural dynamics. This is the different point from the chaotic search which we have proposed in Ref.[7]. Due to this effect, bad exchanges are frequently executed, which means that the state can escape from undesirable local minima. We show that our proposed method enables a wide range search so that it succeeded to improve the performance.

2. QAP

The QAP is one of the most difficult NP-hard combinatorial optimization problems. The QAP is formulated as follows: when two \( n \times n \) matrices, a distance matrix \( D \) and a flow matrix \( R \), are given, we are asked to find an assignment \( p = \{p(1), p(2), \ldots, p(n)\} \) that minimizes an objective function. The objective function of QAP is then defined by Eq.(1):

\[
F(p) = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij} R_{p(i)p(j)}, \tag{1}
\]

where \( p(i) \) is the element \( i \) of the permutation \( p \). If \( p(i) = j \), the element \( i \) is assigned to the location \( j \). In the following, we introduced a 2-exchange method which is a basic algorithm for solving QAPs.

Step1 : A random solution (assignment) \( q \) is made.

Step2 : The objective function \( F(q) \) is calculated.

Step3 : From all the elements, two elements \( s_1 \) and \( s_2 \) are chosen. Then, locations assigned to \( s_1 \) and \( s_2 \) are changed. Let us describe a provided solution as \( q' \).

Step4 : The objective function \( F(q') \) is calculated.

Step5 : If a solution is improved, or \( F(q) > F(q') \), then let \( q = q' \). Return to Step3. When a solution was not improved, even if any two elements \( s_1 \) and \( s_2 \) were chosen, we stop a solution search.

Generally, the 2-exchange method is trapped into undesirable local minima. In this paper, we used a chaotic neural dynamics to escape from the local minima.
3. Proposed method

To control the 2-exchange method by the chaotic neural dynamics\cite{8}, we used Eqs. (2) - (5).

\[
\begin{align*}
\xi_j(t+1) &= \beta \Delta_j(t), \\
\eta_j(t+1) &= w \sum_{i} x_i(t) + w, \\
\zeta_j(t+1) &= k \zeta_j(t) - x_j(t) + (1 - k)\theta,
\end{align*}
\]

where \(\xi_j(t)\) is an external input to the chaotic neuron \(j\), \(\eta_j(t)\) is a feedback input to the chaotic neuron \(j\) from other neurons in the network, \(\zeta_j(t)\) is a refractoriness term of the chaotic neuron \(j\), and \(\Delta_j(t)\) is a gain of the objective function when we change \(p(j)\) to \(p(i)\) by the 2-exchange method, \(k\) is a decay constant, \(w\) is a connection weight of chaotic neurons, \(\alpha\) is a scaling parameter of refractoriness effect, \(\theta\) is a threshold of the chaotic neuron, \(x_j(t)\) is the output of the chaotic neuron \(j\) at time \(t\) and \(f\) is a sigmoidal function defined by \(f(y) = 1/(1 + e^{-\gamma y})\).

In the method in Ref.\cite{6}, when the problem size is \(n\), the \(n \times n\) chaotic neurons are prepared to represent an assignment of \(i\) and \(j\). Namely, if the chaotic neuron \((i, j)\) fires, the element \(i\) is assigned to the location \(j\). Although this method\cite{6} shows good performance, this method uses many neurons. On the other hand, in this study, we use \(n\) chaotic neurons for solving the problem of size \(n\). For this reason, we can reduce the number of neurons. If the chaotic neuron \(i\) fires, we perform the 2-exchange method for the element \(i\). We explain our algorithm as follows.

Step1 : Let \(i = 1\).

Step2 : Internal state values of all chaotic neurons except the chaotic neuron \(i\) are updated asynchronously by Eqs. (2) - (4).

Step3 : The output of all the chaotic neurons except for the chaotic neuron \(i\) are calculated by Eq. (5).

Step4 : If \(\Delta_j(t) < 0\), the elements \(p(i)\) and \(p(j)\) are exchanged by the 2-exchange method and go to Step6. Otherwise go to Step5.

Step5 : If \(\text{max} \{x_j(t+1)\} > 1/2\), the chaotic neuron \(j\) fires and the element \(p(i)\) and \(p(j)\) are really exchanged by the 2-exchange method.

Step6 : If \(i = n\), this iteration is finished. Otherwise let \(i = i + 1\) and return to Step2.

There are two different points from Ref.\cite{7}. First we changed the input term. If we have a bad exchange, we applied a strong input to the dynamics. Namely, in Ref.\cite{7}, \(\Delta_j(t)\) is defined as \(\Delta_j(t) = D_0(t) - D_j(t)\), where \(D_0(t)\) is the present value of the objective function at time \(t\), and \(D_j(t)\) is the value of the objective function that is made by the exchange of \(p(i)\) and \(p(j)\). On the other hand, in this paper, we defined \(\Delta_j(t)\) as \(\Delta_j(t) = D_j(t) - D_0(t)\). Second, we added Step4 to introduce the steepest descent dynamics. Without Step4, only bad exchange is adopted. Then, we cannot get good solutions such as local minima. Therefore, we added Step4.

<table>
<thead>
<tr>
<th>Problem</th>
<th>(i) CS with (n^n) neurons\cite{6}</th>
<th>(ii) CS with (n) neurons\cite{7}</th>
<th>(iii) Proposed method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>gap</td>
<td>gap</td>
<td>gap</td>
</tr>
<tr>
<td></td>
<td>variance</td>
<td>variance</td>
<td></td>
</tr>
<tr>
<td>Bur26a</td>
<td>0.159</td>
<td>0.293</td>
<td>\textbf{0.122 (0.12)}</td>
</tr>
<tr>
<td>Bur26b</td>
<td>\textbf{0.0814}</td>
<td>0.111</td>
<td>0.145 (0.40)</td>
</tr>
<tr>
<td>Bur26c</td>
<td>0.0496</td>
<td>0.130</td>
<td>\textbf{0.0106 (0.13)}</td>
</tr>
<tr>
<td>Bur26d</td>
<td>0.0234</td>
<td>0.080</td>
<td>\textbf{0.00662 (0.13)</td>
</tr>
<tr>
<td>Ste36a</td>
<td>5.65</td>
<td>\textbf{3.86}</td>
<td>4.02 (0.06)</td>
</tr>
<tr>
<td>Ste36b</td>
<td>12.7</td>
<td>8.29</td>
<td>\textbf{4.14 (0.08)}</td>
</tr>
<tr>
<td>Ste36c</td>
<td>4.40</td>
<td>3.68</td>
<td>\textbf{2.21 (0.06)}</td>
</tr>
<tr>
<td>Tai20b</td>
<td>1.80</td>
<td>3.14</td>
<td>\textbf{0.88 (0.04)}</td>
</tr>
<tr>
<td>Tai30b</td>
<td>2.33</td>
<td>1.91</td>
<td>\textbf{0.79 (0.06)}</td>
</tr>
<tr>
<td>Tai40b</td>
<td>3.70</td>
<td>4.58</td>
<td>\textbf{0.99 (0.02)}</td>
</tr>
<tr>
<td>Tai50b</td>
<td>2.21</td>
<td>3.96</td>
<td>\textbf{0.74 (0.02)}</td>
</tr>
<tr>
<td>Tai60b</td>
<td>2.52</td>
<td>2.48</td>
<td>\textbf{0.58 (0.02)}</td>
</tr>
<tr>
<td>Tai80b</td>
<td>2.88</td>
<td>2.08</td>
<td>\textbf{1.03 (0.02)}</td>
</tr>
</tbody>
</table>
4. Result

We evaluated the performance of the proposed algorithm using benchmark problems from QAPLIB[1]. To evaluate the performance, we used the gap which is defined by the following Eq.(6).

$$\text{gap}[_{\%}] = \frac{\text{found best solution} - \text{optimal solution}}{\text{optimal solution}} \times 100.$$  \hspace{1cm} (6)

In this study, we used the following parameter values: $w = 0.0005$, $k = 0.5^\frac{1}{4}$, $\alpha = 1$, $\theta = 0.05$ and $\epsilon = 0.002$. We calculated 100 trials for each parameter $\beta$, and calculated the average gaps across trials. In the numerical experiments, $\Delta_i^j(t)$ is normalized by $d_M r_M$ where $d_M = \max_{ij} d_{ij}$ and $r_M = \max_{ij} r_{ij}$. $d_{ij}$ and $r_{ij}$ are elements of distance matrix $D$ and flow matrix $F$.

Table 1 shows the obtained best gaps for various parameter values of $\beta$. Numerals with bold faced types indicate the best gap, and italic faced types indicate the second best gap. The best parameter $\beta$ for each problem is shown in parentheses. The third row of Table 1 shows the results with the chaotic search(CS) with $n$ neurons in Ref.[7]. The fifth row of Table 1 shows variances of 100 trials for the best. From Table 1, even though our method controls the 2-exchange method with small number of neurons, we can get better performance compared to the method with CS with $n^2$ neurons in Ref.[6] and the method with CS with $n$ neurons in Ref.[7].

To compare the performance of the method, we changed the value of $\beta$ and evaluated the performance. Figure 1 shows the result of the average gap for each $\beta$. In Fig.1, the ordinates show the gap and the abscissas show the values of the parameter $\beta$. The results of the method in Ref.[6] is shown by green lines, the method in Ref.[7] is shown by blue lines and the proposed method is shown by red lines. These figures show that the proposed method can get better performance than the methods in Refs.[6, 7]. In particular, in Bur26a and Tai60b, our proposed method can get better performance than that of the methods in Refs.[6, 7] for a wide range of $\beta$. In Fig.1 shows the average performance. Next, we investigate the variance of the performance.

To investigate the variance of the obtained gaps, we show the best gap of one trial in Fig.2. In Fig.2, ordinates show the gap and abscissas show the number of execution times of the 2-exchange method. The result of the method in Ref.[6] is shown by green lines, the method in Ref.[7] is shown by blue lines and the proposed method is shown by red points. These figures show that almost all the red points exist below the green and blue lines. This means that if we select a suitable parameter $\beta$, we can get better performance than that of the method in Refs.[6, 7] with high probability.
5. Conclusion

In this paper, we proposed a new algorithm for solving Quadratic Assignment Problem (QAP) by chaotic neural dynamics. Concretely, we changed the effect of external input: the worse the exchange is, the stronger the input is applied. This effect enables a wide range search. By comparing with the performance of the method of chaotic search with $n^2$ neurons in Ref.[6] and the chaotic search with $n$ neurons in Ref.[7], our method can get better performance. However, the performance of the proposed method largely depends on value of the parameter $\beta$. Therefore it is an important future work to construct a parameter tuning method.

Acknowledgment

The authors would like to thank Mr. S. Ishii and AGS Corporation for their kind encouragements on this research. The research is supported by Grant-in-Aid (No.26880020, No.15K12137, No.24650116, No.15KT0112) from JSPS.

References


Speeding up of the Traffic Congestion Mitigation by Stochastic Optimization in Deep Learning

Shinnnosuke NAKAMURA†1 Takumi UEMURA12 Gou KOUTAKI13 Keiichi UCHIMURA13

†1 Graduate School of Engineering, Sojo University
†2 Faculty of the Computer Information Science, Sojo University
†3 Graduate School of Science and Technology, Kumamoto University

Email: g15111M03@m.sojo-u.ac.jp, t_uemura@cis.sojo-u.ac.jp, koutaki@cs.kumamoto-u.ac.jp, uchimura@cs.kumamoto-u.ac.jp

Abstract: In recent years, many researchers have become interested in methods for mitigating traffic congestion by optimizing traffic signal parameters. To mitigate traffic congestion over a widespread area, a method using an advanced genetic algorithm and a traffic simulator has been proposed (Nishihara, T., et al., “The Verification with Real-World Road Network on Optimization of Traffic Signal Parameters using Multi-Element Genetic Algorithms”, ITS World Congress, 2012). However, this method consumes considerable time when simulating traffic flow. This paper proposes a method that reduces the processing time of the simulator by using a neural network.

1. Introduction

In recent years, traffic congestion has caused several economic and environmental issues. The Ministry of Land, Infrastructure, and Transportation (MLIT) in Japan has estimated the economic loss caused by congestion to be 12 trillion yen [1]. In addition, it causes a loss of 30 h per year for each person. Moreover, it generates environmental pollution because of the greenhouse gases that are emitted from idling cars. Therefore, the mitigation of traffic congestion has become significant in recent years.

Nishihara et al. [2] addressed this problem by proposing a method that uses both a multiple-element genetic algorithm (ME-GA) and a traffic simulator. However, this method is slow when simulating a road network such as the one shown in Figure 1.

To shorten the processing time, we propose a method that replaces the simulator with a pre-trained neural network (NN). With this replacement, the prediction accuracy has a large effect on mitigation performance. Therefore, we propose a machine learning method that uses deep learning to train the NN.

2. Previous methods

The GreenWave method is an optimization method for traffic signals that is operated in actual service [3, 4]. GreenWave optimizes the traffic signals so as not to stop cars in a certain road section. However, it causes traffic congestion at the interface of this road section.

Xu et al. proposed the GreenSwirl method, which improves on GreenWave. GreenSwirl applies GreenWave to a loop road, and uses a path finding algorithm to allow cars to travel through the loop road optimally. GreenSwirl achieves better mitigation of traffic congestion than GreenWave [5]. However, the design of the loop road and the settings of the road signals are made manually, and thus, this method faces several problems when applied to widespread areas.

Nishihara et al. proposed a method that is constructed from a genetic algorithm (GA) and an evaluation to optimize the traffic signal parameters over a widespread area. Figure 2 shows the processing flow in Nishihara et al.’s method [2]. This method uses the optimization algorithm of multidimensional parameters to optimize the traffic signal parameters over a widespread area. Nishihara et al. used ME-GA to optimize multidimensional parameters in order to obtain generations with better genes. This method comprises the following five steps.
Step 1: Set the actual road network and road signal parameters in the traffic flow simulator.

Step 2: Generate the initial individuals in ME-GA. ME-GA is initialized from the traffic congestion evaluation values derived from the traffic flow simulator.

Step 3: Derive better road signal parameters by calculating new generations in ME-GA.

Step 4: Derive the traffic congestion evaluation values of each road signal parameter from the traffic flow simulator.

Step 5: Apply each traffic congestion evaluation value to the ME-GA and iterate steps 3 to 5 until a certain number of generations has been calculated.

This method iterates the operation of the traffic flow simulator to optimize the traffic signal parameters; therefore, the processing time of the traffic flow simulator significantly affects the total time of the system. Nishihara et al.’s method consumes 19 hours to optimize the traffic signal parameters in the simple road network shown in Figure 1. The processing time of the traffic flow simulator needs to be shortened in order for this method to be operated in actual service.

3. Proposed method

3.1. Outline

In this paper, to shorten the processing time, we propose a traffic signal optimization method that replaces the traffic simulator with a pre-trained NN, in order to predict with high accuracy using deep learning.

The learning methods of deep learning can be classified into the following four types: autoencoder (AE) [6], restricted Boltzmann machine [7, 8], convolutional NN [9-11], and recurrent NN [12, 13]. Among them, we choose AE because it is the most versatile and has the same structure as that of a general NN. In addition, we use a stack denoising AE (SDA) because it is generally regarded as the method that obtains the best performance for AE.

We use a stochastic optimization approach, called Adam, to optimize the network parameters for deep learning. In recent years, stochastic optimization has become a major method in deep learning that accelerates learning convergence and reduces generalization error. Adam is regarded as learning more quickly than other conventional parameter optimization methods such as AdaGrad or RMSProp [14].

Figure 3 shows the process flow of the proposed method, which comprises two processes. The first is the learning process for predicting the output value of the traffic flow simulator. This process samples the learning data from the traffic flow simulator and extracts the features for the NN from the data via unsupervised learning using SDA. After feature extraction, the NN predicts the output value via supervised learning using the back propagation method. The second process optimizes the traffic signal parameters via a trained NN and ME-GA. ME-GA better predicts the traffic signal parameters by using the congestion evaluation values that are outputted from the NN. The traffic signal parameters are optimized by iterating the optimization process.

3.2. Properties of NN

Supervised learning is performed after feature extraction by deep learning using the back propagation method. In this study, the training data are normalized as a real number from [0, 1]. We use a sigmoid function for the activation function, as shown in the following equation.

\[
sigmoid(z) = \frac{1}{1+e^{-z}}
\]  

Here, \( Z \) is the output value of the element and \( \alpha \) is the gain.

The error value \( E \) in the output layer is derived from the mean square error, as shown in following equation.

\[
E = \frac{1}{2} \sum_k (T_k - O_k)^2
\]

Here, \( T_k \) shows the average value of the training data and \( O_k \) shows the \( k \) th sample value of the output.

Table 1 shows the Adam parameter optimization procedure. Parameter \( \theta \) is the parameter to be determined,
\( f \) is the objective function, \( g \) is a gradient derived from \( f \), and \( t \) is the number of iterations for learning. In addition, the recommended parameters for Adam are shown at the top of Table 1.

4. Experiment

We carried out two experiments to confirm the effectiveness of the proposed method. Experiment 1 compares the congestion mitigation performances of the proposed method, Nishihara et al.'s method, and the actual measured traffic signal parameters. Experiment 2 evaluates the calculation speed of the proposed method and Nishihara et al.'s method.

4.1. Experimental setup

Table 2 shows the specifications of the computer used for each method.

4.1.1. Properties of the traffic flow simulator

The traffic flow simulator used in the evaluation is Aimsun 6.1. The road network in these experiments is the same as that used to evaluate Nishihara et al.'s method (shown in Figure 1) [2]. In addition, the traffic volume and agent of each vehicle in Aimsun 6.1 are set the same as those for Nishihara et al.'s method.

The traffic signal parameters consist of Cycle, Split, and Offset signals. These are set for each intersection. Aimsun 6.1 outputs WaitOut, Inside, and GoneOut data. WaitOut is the number of vehicles that cannot enter the road network because of traffic jams and are located outside of the road network to be optimized. Further, Inside is the number of vehicles in the road network to be optimized. Finally, GoneOut is the number of vehicles that have left the road network to be optimized.

4.1.2. Properties of the NN

In the deep learning for these experiments, the NN pre-learns using SDA before it is trained using error back propagation. The NN does not include an output layer because of the unsupervised learning in pre-learning.

We construct an NN for every congestion evaluation value because the congestion evaluation values do not affect each other.

Mini-batch learning is used as the learning method. To consider versatility, each number of iterations for learning is set to minimize the differences between generalization and training errors in the learning process.

4.1.3. Properties of ME-GA

The ME-GA parameters are set such that the number of generations is 500 and the population is 300. The other settings are the same as those in Nishihara et al.'s method. (The number of generations and population of Nishihara et al.'s method are 100 and 75, respectively). The evaluation value \( F \) for ME-GA is derived from the following two equations.

\[
F = \exp \left( \frac{V_{\text{wo}}}{C_w} \right) + \exp \left( \frac{V_{\text{in}}}{C_i} \right) + \exp \left( \frac{C_d}{C_d} \right) \tag{3}
\]

\[
C = \frac{t_{\text{delay}}}{TTD} \tag{4}
\]

Here, \( V_{\text{wo}} \) is WaitOut and \( V_{\text{in}} \) is Inside. Further, \( t_{\text{delay}} \) (DelayTime) is the average difference between the actual running time and the ideal running time; the total travel distance (\( TTD \)) indicates the sum of the travel distance of all vehicles in the simulation; and \( C_w, C_i, \) and \( C_d \) are weights for the congestion evaluation value, where \( C_w = 100, C_i = 500, \) and \( C_d = 500 \).

4.2. Experimental results

Table 4 shows the results of the first experiment. Here, low values for WaitOut and Inside indicate that the traffic situation is good. A large value for GoneOut also indicates that the traffic situation is good.

The results of this experiment confirm that the traffic congestion resulting from Nishihara et al.'s method and the proposed method are equal. The results also show that the proposed method performs worse than Nishihara et al.'s method when the number of generations and population are set to be the same as those in Nishihara et al.'s method. Therefore, the method for predicting the output values of the traffic flow simulator would probably be improved by reconsidering the samples used for pre-training.

Table 5 shows the results of Experiment 2. The processing time in the proposed method is about 15 minutes, while that in Nishihara et al.'s method is about 19 hours. Therefore our proposed method achieves a processing time that is 99% shorter than that in Nishihara's method.

### Table 1 Algorithm of Adam

| \( m_0 \) | 0 |
| \( v_0 \) | 0 |
| \( t \) | 0 |

while learn number do

\( t \leftarrow t + 1 \)

\( g_t \leftarrow \nabla_{\theta_{t-1}} \)

\( m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t \)

\( v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \)

\( \tilde{m}_t \leftarrow m_t / (1 - \beta_1^t) \)

\( \tilde{v}_t \leftarrow v_t / (1 - \beta_2^t) \)

\( \theta_t \leftarrow \theta_{t-1} - \alpha \cdot \tilde{m}_t / (\sqrt{\tilde{v}_t} + \varepsilon) \)

end while
5. Conclusion

In this study, to shorten the processing time while maintaining accuracy, we proposed a traffic signal parameter optimization method that uses a pre-trained NN instead of a traffic flow simulator.

The experimental results show that we achieved a similar level of mitigation for traffic congestion and a 99% shorter processing time than Nishihara et al.’s method. However, our method performed worse than theirs when the numbers of generations and individuals were set to the same values as in their method. In addition, the proposed method could optimize the traffic signal parameters in 15 min. However, VICS updates information every 5 min generally [15]. Therefore, the processing time of the system should be shortened to less than 5 min in order to be practical for actual service.

In future work, we plan to apply parallel processing to the calculation of individual evaluations in ME-GA or increase the sampling data in deep learning.

**References**


Dynamic Behavior of a Bouncing Ball

Shu Karube†, Takuji Kousaka†† and Yuya Kawazu†††

†National Institute of Technology, Oita College, 1666 Maki, Oita, 870-0152, Japan
††Oita University, Faculty of Engineering, 700 Dannoharu, Oita, 870-1192, Japan
†††Mechanical-Environmental System Engineering Major, National Institute of Technology, Oita College,
1666 Maki, Oita, 870-0152, Japan
Email: karube@oita-ct.ac.jp, takuji@oita-u.ac.jp, ame15084@ga.oita-ct.ac.jp

Abstract– Bouncing ball system is often used for describing collision dynamics of various mechanical systems. We produced the experimental bouncing ball equipment using a ping-pong ball, a table tennis paddle with rubber and a shaker for vibrating the paddle. In this study, we investigate frequency responses of the bouncing ball system under constant paddle amplitude condition and constant paddle acceleration amplitude condition experimentally, and compare our experimental results with simulations. Under constant amplitude condition, we found that the maximum height of the ball increases stepwise as varying the paddle frequency by both experiment and simulation. Under constant acceleration amplitude condition, we found the maximum height of the ball decreases exponentially as varying the vibrated plate frequency by both experiment and simulation. Summarizing the above, we observed the nonlinear characteristics that the height of ball increases stepwise in simple bouncing ball system.

1. Introduction

Generally, most mechanical systems have gaps, such as a pair of gears with backlash. In these systems, nonlinear vibrations by collision phenomenon often occur. A bouncing ball system is effective to explain these nonlinear vibrations on collision systems.

The bouncing ball system can be applied to various mechanical systems. For example, M. Paskota analyzed and controlled collision vibration of a molding box during in aluminum production [1]. K.Sakai et al. analyzed dynamics of the tractor on the rigid road using a bouncing ball model [2]. As experimental studies, A.Kini et al. studied the impact vibration when a super ball collided with a fixed table with rubber [3].

We produced the bouncing ball equipment using a ball and a paddle for table tennis [4]. Our equipment had the advantage that frictional resistance for the ball was small. However, the vibration direction of the paddle caused a slight error, hence the unbalanced support of the paddle. The paddle is fixed to a shaker at its handle and vibrated. This way distorted the vertical direction of the surface of the racket collided with the ball.

In this study, we improve our previous bouncing ball equipment. The surface of paddle is fixed at the center of shaker, in order to keep precise vertical vibration of the paddle. Next, we carry out the frequency response experiments under both constant amplitude condition and constant acceleration amplitude condition, then we compare these experimental results with our simulation results.

2. Experimental setup

Fig.1 shows the photograph of our experimental setup we produced. Fig.2 shows the schematic illustration of the equipment. The device consists of four parts mainly, i.e. a ping-pong ball, a paddle for table tennis, a ball guide and a shaker. The shaker vibrates the paddle periodically to the vertical direction. The ball is vibrated by the paddle, and jumps on the paddle. The displacement of the ball is measured by a sensor 1, which is a laser displacement sensor LK-G405 produced by Keyence corporation. The ball guide consists of a pair of top and bottom plates and four stainless steel poles fixed between the plates. The ball guide limits the horizontal movement of the ball by the method surrounding with four stainless steel poles whose diameter is 3mm in order to keep the ball at the same position of the paddle. There is a hole to let the ball go through, whose diameter is 40mm, at the center of the top plate and the bottom plate. The four stainless steel poles are set around the hole at intervals of 90degrees. The distance between the top and the bottom plate is 300mm, they are supported with four support pillars whose length are 767mm. The displacement of the paddle is measured by a sensor 2, which is a laser displacement sensor LK-G155 produced by Keyence corporation. The positive direction of the sensors is upward. When the ball stops, the distance of the ball and the sensor 1 is 400mm. The mass of the ball is each 2.8g. The diameter of the ball is 39.8mm.

Fig.3 shows the paddle vibration control system. The system is feedback controlled by vibration controller K2 sprint, which is produced by IMV corporation. The vibration controller is able to control the acceleration, velocity, and amplitude of the paddle by using the signal from the acceleration pickup fixed on the paddle.
3. Simulation

3.1. Bouncing ball model

Fig.4 shows physical model of the bouncing ball system that we use in this study. \( m_1 \) denotes the mass of the ball. \( x_1 \) and \( x_2 \) denote each the displacement of the ball and the paddle. The motion equations are defined by Eq.1.

\[
\begin{align*}
    m_1 \ddot{x}_1 &= -m_1 g \\
    \dot{x}_2 &= a_0 \sin \omega t
\end{align*}
\]

(1)

The parameters in Eq.1 are shown as follows: \( g \) is the gravitational acceleration, \( \omega \) is the angle frequency of the paddle, \( a_0 \) is amplitude of the paddle. When \( x_i = x_1 \) the ball collides with the paddle. Suppose \( m_i >> m_1 \), where \( m_2 \) is the mass of the paddle. Therefore, the change of the velocity of the ball \( \dot{x}_i \) is given by

\[
\dot{x}_i^+ = -e \dot{x}_i^- + (1 + e) \dot{x}_2
\]

(2)

where \( \dot{x}_i^- \) and \( \dot{x}_i^+ \) is the velocity of the ball before and after impact.

The parameters used in the simulation are as follows: \( e = 0.68 \), \( m_1 = 2.8 \) g, \( a_0 = 0.69 \) mm, \( g = 9.8 \) m/s².

3.2. Simulation condition

In this paper, we use Runge-Kutta-Gill 4th order Method with time step (128 \( \dot{f} \))

\( T \) transient number 100,000 and data number 100,000.

The velocity and acceleration of the paddle are defined by Eq.3 and 4.

\begin{align*}
    \ddot{x}_2 &= a_0 \omega \cos \omega t \\
    \ddot{x}_1 &= -a_0 \omega^2 \sin \omega t
\end{align*}

(3) (4)

By defined the acceleration amplitude \( A_0 = \sqrt{\dot{x}_2^2} \) max as Eq.4 can be rearranged to Eq.5.

\[
A_0 = a_0 \omega^2
\]

(5)
4. Experimental and Simulation result

In this section, we investigate frequency responses of our bouncing ball system. Increasing frequency of the paddle $f$ from 20Hz to 50Hz by 1Hz, we measure the maximum value of the displacement of the ball $x_{1\text{max}}$ in each frequency. We carry out two types of the frequency response experiment. One is the experiment while keeping the paddle amplitude $a_0$ uniformity, which is called as constant amplitude condition. The other is this while keeping the acceleration amplitude $A_0$ uniformity, which is called as constant acceleration. We determine $a_0 = 0.69\text{mm}$ under constant amplitude condition, $A_0 = 35\text{m/s}^2$ under constant acceleration.

Fig.5(a) shows experimental and simulation results under constant amplitude condition. The vertical axis indicates the maximum displacement of the ball $x_{1\text{max}}$, and the horizontal axis indicates the frequency of the paddle $f$. Both experimental and simulation result, the graphs show stepwise increase of $x_{1\text{max}}$ as varying the paddle frequency $f$.

Fig.5(b) shows experimental and simulation result under acceleration displacement constant. Both experimental and simulation result, the graphs show exponential decrease of $x_{1\text{max}}$ as varying the paddle frequency $f$. The trend of increasing $x_{1\text{max}}$ under constant amplitude condition in Fig.5(a) can be explained by Eq.6.

$$\dot{x}_2|_{a_0=\text{const.}} = a_0\omega \propto f$$ (6)

Eq.6 indicates that maximum velocity of the paddle $|\dot{x}_2|_{\text{max}}$ is proportional to $f$. Similarly, the trend of
decreasing \( x_{1,\text{max}} \) under constant acceleration amplitude condition in Fig.5(b) can be explained by Eq.7

\[
\left[ x_{2,\text{max}} \right]_{A_0=\text{const.}} = A_0 \omega = A_0 \omega \propto 1/f
\] (7)

Eq.7 indicates that \( [x_{2,\text{max}}] \) is inversely proportional to \( f \). Bouncing height of the ball depends on the velocity of the paddle when the ball colliding the paddle. Hence, under constant amplitude condition, according to Eq.6, \( x_{1,\text{max}} \) monotonic increases, and constant acceleration amplitude condition, according to Eq.7, \( x_{1,\text{max}} \) monotonic decreases. However the stepwise increasing of \( x_{1,\text{max}} \) under constant amplitude condition in Fig.5(a) cannot be explained by Eq.6. We suppose that the trend of stepwise increasing is caused by nonlinear characteristics in our simple bouncing ball system.

5. Poincare map in constant amplitude

In this section, we detailed simulation for previous results in Fig.5(a).

Fig.6 is the bifurcation diagram under constant amplitude condition, in the case when the paddle frequency \( f \) is increased from 20Hz to 50Hz by 0.05Hz. In Fig.6, the trend of stepwise increase of \( x_{1,\text{max}} \) is clearer than the graph of the simulation result in Fig.5(a). We suppose that the bifurcations are occurred at the intermittent points like stepwise. Next, we investigate the dynamics of our system when the frequency is before and after bifurcation, \( f = 28\text{Hz and 32Hz} \).

Fig.7 shows Poincare maps under constant amplitude condition shown by Fig.6. Fig.7(a) and (b) show the Poincare map at 28Hz and 32Hz, i.e. before and after bifurcation. The Poincare map in Fig.7(a) shows three regions of Poincare plots, A, B and C. In contrast, Fig.7(b) shows four regions, A, B, C, and D. The region D is occurred by the bifurcation. Therefore, we suppose that there are the bifurcations which keep the dynamics and add new region.

6. Summary

In this study, we investigated the dynamics of a bouncing ball system, which was consisted of a ball and a paddle for table tennis, experimentally and numerically. The paddle was vibrated by a shaker. First, we improved our previous bouncing ball equipment and was capable of more precision experiments. Second, we carried out frequency response experiments under two types of condition. One was the constant amplitude condition, which was keeping the amplitude of the paddle as varying the frequency of the paddle by vibration control system. The other condition was the constant acceleration amplitude condition, which was keeping the acceleration amplitude of the paddle. Under the constant amplitude condition, our experiment showed the maximum height of ball increased stepwise. Same trend was shown by our numerical simulation. Calculated Poincare maps showed occurrence of new region of Poincare plots surrounding stepwise increasing points.

As a result, we supposed that there was the bifurcations which kept the conventional structure of the dynamics and added new structure in the simple bouncing ball system.

7. Acknowledgement

This work was supported by JSPS KAKENHI Grant Numbers JP16H02876.

References

Slip Compensation of Induction Motor using Sensorless Control

Tomoyuki Dansako†, Nobuyuki Kasa‡, Hiroyuki Asahara‡

†Graduate School, Okayama University of Science
‡Faculty of Engineering, Okayama University of Science

1-1 Ridai-cho, Kita-ku, Okayama, 700-0005 Japan
Email: t15em03dt@ous.jp, kasa@ee.ous.ac.jp, asahara@ee.ous.ac.jp

Abstract—An induction motor is a kind of AC motor. It is known that this motor is rotated later than synchronous speed, by applying a load. Therefore, the present study is estimated rotational speed by using an observer, and compensates the slip of an induction motor in sensorless control in order to control the rotational speed against the synchronous speed.

1. Introduction

To describe the operating principles of the squirrel-cage induction motor. Figure 1 shows the structure of the motor. When current is applied to the stator, the induced current is flowed through the rotor. Then, it becomes an electromagnet, and the mechanism that rotates according to the rotating magnetic field created by the stator. The structure of this motor is also simple, and low cost because it is not using the precious metal. In addition, it can be driven by a maintenance-free because the commutator is not exist.

There is a phenomenon of slip in an induction motor. The slip is difference between the actual rotor and synchronous speed. This motor rotates slightly delayed by the slip. When the load is applied, the torque corresponding to the load is increase. And the slip is increase by torque, the rotational speed is reduced. Accordingly, the purpose of present study is to compensate for the slip using the sensorless control.

Moreover, in the present study, driving an induction motor using the model equation of the synchronous motor as a new method. In Ref [2], being described that the induction motor and synchronous motor can be adapted to a common. By the model equation in common, it can be controlled with a small amount of the program, such as the parameter measurement and the sensorless control. There is an advantage that can be more easily driven.

2. Estimation Method

2.1 α – β Coordinate Transformation

It is extremely difficult to control the induction motor in the state of the three-phase AC. Thereby, perform a-b transform to convert from the three-coordinate system to the two-coordinate system, and to simplify the control of the motor. The three-phase AC has a property to become zero when add to three sine waves. By use of this property, it is possible to handle the three-phase AC as an equivalent two-phase AC.

Let the three-phase current be \(i_u, i_v, i_w\) and the two-phase transformation of the current be \(i_α, i_β\), represented in the following equation.

\[
i_u + i_v + i_w = 0 \quad i_α = i_u, \quad i_β = \frac{(i_u + 2i_w)}{\sqrt{3}}
\]

2.2 d – q Coordinate Transformation

Rotating coordinate system can handle current as a DC. Thus, by converting from the fixed coordinate to the rotating coordinate, analysis and control becomes very easy. Moreover, the rotating coordinate system can be divided current into two components, handle q-axis as the torque current component and the d-axis as the field current component. And it is possible to control them independently.
2.3 Speed Sensorless Control Method

As estimation method, assemble the controlled object of the system on a microcomputer. Then, give the same input to the actual and assembles system, and compare outputs. By adjusting to eliminate the error of the output, the system on the microcomputer and controlled system becomes almost same parameter. Thus, the information of the motor can be obtained.

$$i_d = i_a \cdot \cos \theta + i_b \cdot \sin \theta$$
$$i_q = -i_a \cdot \sin \theta + i_b \cdot \cos \theta$$

\[ P \begin{bmatrix} f_d \\ f_q \end{bmatrix} = \begin{bmatrix} \frac{\mu L_d}{L_q} & 0 & 0 & -\frac{\mu L_q}{L_d} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} i_d \\ i_q \end{bmatrix} + \begin{bmatrix} \varepsilon_d \\ \varepsilon_q \end{bmatrix} \quad \cdots (1) \]

The previous equation is input. Create a feedback system to input \(-w\) and output \(v\), and set the vector to be stable.

\[ v = \begin{bmatrix} i_q \\ i_q \end{bmatrix} \frac{E_d}{E_q} \cdots (3) \]
\[ -w = \begin{bmatrix} i_d \\ i_q \end{bmatrix} \begin{bmatrix} 0 & \Delta \omega \\ -\Delta \omega & 0 \end{bmatrix} \begin{bmatrix} i_d \\ i_q \end{bmatrix} \cdots (4) \]

\[ \Delta \omega \text{ is obtained by the transfer function } G = \frac{-v}{w} = 1, \text{ the estimated speed value is obtained by inputting the PI compensator.} \]

\[ \hat{\omega} = K_p \Delta \omega + K_i \int \Delta \omega dt \cdots (5) \]

3. Experimental Results

3.1 Experimental Systems

An inverter consists of the hall effect sensors, an intelligent power module, a microcomputer which is RENESAS RX62T. The specifications of induction motors are power output 0.75 kW and 1.5kW, 4-poles, MITSUBISHI squirrel-cage induction motor.

3.2 Experimental results

The actual q-axis current is followed by the q-axis command value. Figure 6 is the graph obtained by adding the load until 0.8Nm to the induction motor.
These results, increased about 2 rpm per 0.5 Nm, the amount of q-axis current raised about 0.07 A. Therefore, the increase of the q-axis command value is divided by 0.035[A], and substituted into the estimated speed value as shown in Figure 7. The result of inputting this value is shown in Figure 8. This graph is the result of torques and rotational speeds.

\[
\begin{align*}
I_{1d} & = \frac{R_2 L_m i_{1d} + L_2 i_{2d}}{L_2} \\
I_{1q} & = \frac{L_1 i_{1d} + R_1 i_{1q} + L_2 i_{2q}}{L_2} \\
I_{2d} & = \frac{-\omega_1 L_2 i_{1d} - \omega_2 L_2 i_{2d}}{L_1} \\
I_{2q} & = \frac{L_1 i_{1q} + R_1 i_{1d} + \omega_2 L_2 i_{2d}}{L_1} \\
\end{align*}
\]

\( v_{1d}, v_{1q}, v_{2d}, v_{2q} \) : Primary and Secondary d-q axis voltage
\( i_{1d}, i_{1q}, i_{2d}, i_{2q} \) : Primary and Secondary d-q axis current
\( R_1, R_2 \) : Primary and Secondary resistances
\( L_1, L_2 \) : Primary and Secondary self-inductance
\( L_m \) : Mutual inductance
\( \omega_s \) : Slip frequency
\( \omega_1 \) : Electrical angle

Depending on the characteristics of the squirrel-cage induction motor, the voltage on the secondary is zero because it is shorted circuit. The flux linkage can be expressed by the following equation.

\[
\varphi_{2d} = L_m i_{1d} + L_2 i_{2d} \quad \text{... (8)}
\]

\[
\varphi_{2q} = L_m i_{1q} + L_2 i_{2q} \quad \text{... (9)}
\]

The following equation is obtained by substituting these equations to the basic equation.

\[
\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -L_m R_2 & 0 & R_2 + p L_2 & -\omega_2 L_2 \\ 0 & -L_m R_2 & \omega_2 L_2 & R_2 + p L_2 \end{bmatrix} \begin{bmatrix} I_{1d} \\ I_{1q} \\ I_{2d} \\ I_{2q} \end{bmatrix} \quad \text{... (10)}
\]

When the direction of the secondary flux linkage \( \varphi_2 \) is coincident with the d-axis, \( \varphi_{2q} \) is zero. Thus, the slip frequency \( \omega_s \) is represented by the following equation.

\[
\omega_s = \frac{R_2 L_m}{L_2 \varphi_2} i_{1d} \quad \text{... (11)}
\]

From this equation, the slip frequency is calculated by the secondary resistance and q-axis current. And, the estimated rotor speed can be compensated by the calculated slip frequency \( \omega_s \).

### 3.3 Slip Compensation Result

When the q-axis current increases the amount of 0.035 A, the rotational speed has decreased 1 rpm. Using the characteristic equation of the induction motor from the results, the slip is compensated using two induction motors. These results are shown in Figures 9, 10.
Fig. 9 Slip compensation result of 0.75kw induction motor.

As a result, compensating for slip in the same manner as when compensated using q-axis current 0.035 A. In addition, the slip of the induction motor of 1.5kW was almost compensated.

Fig. 10 Slip compensation result of 1.5kw induction motor.

4. Conclusion

In this study, it was possible to compensate for slip from the q-axis current command of the induction motor using a sensorless control. Moreover, using a characteristic equation of the induction motor, it was able to compensate to the slip by calculating the slip frequency from the increase in the q-axis current. Future issues are examining the rotational speed when applied a greater load than 0.8 Nm.

References


Nonlinear Oscillation of a Photovoltaic Cell Booster

Yushi Hasesaka†, Keisuke Mori‡, Takuji Kousaka‡, Hirokazu Ohtagaki‡ and Hiroyuki Asahara‡

†Faculty of Engineering, Okayama University of Science, 1-1 Ridai-cho, Kita-ku, Okayama-shi, Okayama 700-0005, Japan.
‡Faculty of Engineering, Oita University, 700 Dannoharu, Oita-shi, Oita 870-1192, Japan.
Email: [yushi,mori]@nonlineargroup.net, takuji@oita-u.ac.jp, {ohtagaki,asahara}@ee.ous.ac.jp

Abstract—In this study, we report fundamental characteristic of a photovoltaic cell booster. We show the circuit model, which is current-controlled DC/DC boost converter fed with photovoltaic module. Using the sampled data of the waveforms at the clock intervals, nonlinear oscillation of the circuit is investigated briefly. The circuit characteristic obtained in this report will help to control and design the photovoltaic cell booster in our future study.

1. Introduction

The clean energy technology is developing in recent years, and many power generation devices and its peripheral circuits are further along in development. The photovoltaic cell booster is a clean energy power generation system, which is a DC/DC converter fed with photovoltaic cell.

Because the photovoltaic cell is regarded as a current source, we often connect it to DC/DC converter, and control the current to track the maximum power point of the photovoltaic cell. There are many equivalent circuits of DC/DC converter fed with photovoltaic cell, and it is known that rich nonlinear phenomena are observed due to the switching events [1, 2]. From the viewpoints of circuit design and circuit theory, clarifying dynamic behavior of DC/DC converters fed with photovoltaic cell is important. However, circuit equation of DC/DC converter fed with photovoltaic cell have two or more dimensional topology and nonlinear characteristic, and its complexity makes the analysis difficult. Therefore, Ref. [3] proposed a simplified model, by assuming the photovoltaic cell to be a current source with piecewise smooth characteristic.

Now, we can design suitable circuit parameter based on tools, which are circuit model and its analysis technique, reported in Refs. [1, 2], and we can understand qualitative property based on the tools reported in Ref. [3]. This paper aims to propose an intermediate circuit model which has one-dimensional topology, and well simulate output characteristic of photovoltaic cell without using piecewise smooth characteristic.

First, we simulate output characteristic of photovoltaic cell based on Ref. [4]. Next, we show circuit dynamics of DC/DC convener fed with photovoltaic cell, and explain switching events. Finally, we discuss characteristics of the proposed circuit, and demonstrate nonlinear oscillation.

2. Photovoltaic Cell Booster

Table 1 shows output characteristics of the photovoltaic cell, which is used in this study. Table 1 shows output characteristics of the photovoltaic cell, which is used in this study. The output of the photovoltaic cell is calculated based on Ref. [4] as shown in Fig. 1.

Figure 2 shows DC/DC boost converter fed with photovoltaic cell, which we call photovoltaic cell booster. The circuit parameters are follows:

\[
L = 1\text{[mH]}, \quad C = 100\text{[μH]}, \quad I_{\text{ref}} = 0.92\text{[A]}, \quad T = 17\text{[μs]}. \tag{1}
\]

If the switch is ON, the circuit equations are given by

\[
L \frac{di}{dt} = V_{\text{pv}} \tag{2}
\]

and

\[
C \frac{dv}{dt} = -\frac{V}{R} \tag{3}
\]

Likewise, if the switch is OFF, the circuit equations are given by

\[
L \frac{di}{dt} = V_{\text{pv}} - v \tag{4}
\]

and

\[
C \frac{dv}{dt} = i - \frac{v}{R}. \tag{5}
\]

Note that \(V_{\text{pv}}\) denotes output voltage from photovoltaic cell.

The drive circuit includes comparator and flip-flop. Using the current transformer, we sense the inductor current (i.e. output current of photovoltaic cell). The output of comparator, which compares inductor current and reference current, is input into flip-flop, and switching signal is generated. Note that we set reference current to the same value of the current at maximum power point of photovoltaic cell (i.e. \(I_{\text{ref}} = I_{\text{pm}}\)).

<table>
<thead>
<tr>
<th>Maximum Power (P_{\text{m}})</th>
<th>14 [W]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current at Maximum Power (I_{\text{pm}})</td>
<td>0.92 [A]</td>
</tr>
<tr>
<td>Voltage at Maximum Power (V_{\text{pm}})</td>
<td>15.3 [V]</td>
</tr>
<tr>
<td>Short Circuit Current (I_{\text{sc}})</td>
<td>1.0 [A]</td>
</tr>
<tr>
<td>Open Circuit Voltage (V_{\text{oc}})</td>
<td>18.6 [V]</td>
</tr>
</tbody>
</table>

Table 1: Photovoltaic cell (BT432S-MRN).
3. Nonlinear oscillation observed in Equivalent circuit

We assume that capacitor voltage is constant value $E_0$. Therefore, the circuit equation is rewritten as follows:

\[
L \frac{di}{dt} = \begin{cases} 
V_{pv}, & \text{for ON} \\
V_{pv} - E_0, & \text{for OFF}
\end{cases}
\]

(6)

The transfer factor is described as

\[
M = \frac{E_0}{V_{pv}} = \frac{1}{D'}
\]

(7)

where $D' = 1 - D$. In addition, output current of photovoltaic cell is described as

\[
I_{pv} = \frac{E_0}{R} = I_{pv}D'
\]

(8)

Therefore, we get

\[
E_0 = \sqrt{I_{pv}V_{pv}R} \approx \sqrt{I_{pm}V_{pm}R}.
\]

(9)

Note that we rewrite $I_{pv}$ and $V_{pv}$ by $I_{pm}$ and $V_{pm}$. In the following, we use dimensionless values: $t = Lt$, for the sake of the simplicity.

Figures 3 and 4 show waveform behavior of the capacitor voltage and inductor current; Fig. 3 is calculated using Eqs. (2)-(5), whereas Fig. 4 is calculated using Eq. (6). It is clear that the same waveforms are observed in the equivalent circuit. Therefore, we analyze nonlinear oscillation of the equivalent circuit in the following analysis.

There are two types of waveform behavior of the inductor current during clock interval. The switch keeps ON in the first case which we call case-1, whereas the switch changes its position from ON to OFF in the other case which we call case-2. We assume an initial value at $\tau = kT$, $k = 1, 2, 3, \cdots$, by $i_k$, and that of at $\tau = (k + 1)T$ by $i_{k+1}$. We define the solution of Eq. (6) as follows:

\[
i(\tau-kT) = \varphi(\tau-kT, i_k, \lambda) = \begin{cases} 
\varphi_1(\tau-kT, i_k, \lambda_1), & \text{for ON} \\
\varphi_2(\tau-kT, i_k, \lambda_2), & \text{for OFF}
\end{cases}
\]

(10)

where $\lambda, \lambda_1$, and $\lambda_2$ are parameters.

In case-1, the $i_{k+1}$ is defined as follows:

\[
i_{k+1} = F(i_k) = \varphi_1(T, i_k, \lambda_1).
\]

(11)

Likewise, the $i_{k+1}$ is defined by following equation in case-2.

\[
i_{k+1} = F(i_k) = P_2(P_1(i_k)),
\]

(12)

where functions $P_1$ and $P_2$ in Eq. (12) are expressed by the following equations.

\[
P_1(i_k) = \varphi_1(\tau_{on}, i_k, \lambda_1) = I_{ref}
\]

(13)
Figure 3: Numerical simulation of photovoltaic cell booster ($R = 75$).

Figure 4: Numerical simulation of equivalent circuit ($R = 75$).

$$P_2(i_0) = \varphi_2(T - \tau_{\text{on}}, P_1(i_k), I_2) \quad (14)$$

Note that $\tau_{\text{on}}$ denotes ON time during which the switch keeps ON, and $i_0$ is an initial value, which satisfies $i_0 = P_1(i_k) = I_{\text{ref}}$.

Figure 5 shows $i_k$ versus $i_{k+1}$ plane which we call return map. We change the resistance by $R = 50[\Omega]$, $R = 80[\Omega]$, $R = 100[\Omega]$, and $R = 175[\Omega]$. We can observe nonlinear phenomena in the circuit. For example, the fixed point is observed at $R = 50[\Omega]$, the period-2 orbit is observed at $R = 80[\Omega]$, the period-4 orbit is observed at $R = 100[\Omega]$, and aperiodic orbit is observed at $R = 175[\Omega]$.

4. Conclusion

We studied DC/DC boost converter fed with photovoltaic cell. First, we showed output characteristic of photovoltaic cell, and then we explained circuit dynamics. Finally, we proposed equivalent circuit, and demonstrated nonlinear phenomena observed in the equivalent circuit by using the return map. In future, we study this circuit in detail.

References


Chaos controlling with clock pulse modulation for DC-DC converter circuit

Daisuke Ito†

†Department of Electronic Systems Engineering, The University of Shiga Prefecture
2500 Hassaka-cho, Hikone-shi, Shiga, 522-8355 JAPAN
Email: ito.d@e.usp.ac.jp

Abstract—The DC-DC converters are one kind of hybrid dynamical systems, and they present typical nonlinear phenomena; bifurcation phenomena and chaotic attractors. On engineering view points, they are undesigned behavior because they cause abrupt motions or noise. To suppress them, various controlling techniques have been proposed. On DC-DC converters, the period of pulse inputs is an important parameter for behavior of the circuit. In this paper, we investigate the chaos controlling method by adjusting the interrupt dynamical events caused by pulse inputs, and propose two-type pulse modulation schemes.

1. Introduction

The systems with interrupt events that change dynamical behavior of them such as switches are treated as hybrid dynamical systems. These features are observed in many engineering’s fields [1]. Due to the non-linearity of interrupt events, rich complex behavior appears, i.e. bifurcation phenomena and chaotic attractors. They are also observed on the one-dimensional piece-wise linear system, and it is confirmed that border-collision bifurcations play an important role [3, 2, 4].

In the electrical circuit, hybrid dynamical systems exist by featuring the electrical switches, i.e. mechanical switches and MOS-FETs. A DC-DC converter circuit is one of hybrid dynamical systems. Some reports issue the bifurcation and chaos, and the influence for its electrical characteristic is investigated [5, 6]. On the view point of performance as converters, chaotic attractors are similar to noise-like responses, and they can be considered as the behavior to avoid. In recent studies, various controlling schemes via chaos controlling are proposed for DC-DC converter models [7, 8, 9]. Kousaka, et.al proposed the controlling scheme with varying the source voltage. We also have proposed the method with varying reference voltages of comparators, but they mean output voltage values, and it should be fixed as objective output voltage values. On the other hand, the pulse width modulation (PWM) input is used to drive the converter circuit. Its duty ratio decides the output voltage, and the period of the pulse have influence the circuit’s behavior mainly. Accordingly, the frequency modulation has possibility to control the circuit, and chaos controlling by perturbation for frequencies.

In this study, we try to suppress chaotic phenomena on the DC-DC converter model circuit by the clock pulse modulation. The parameter perturbation based on the feedback control is used. The controller gain is designed with a pole assignment method. At first, we explain the design procedure, and show the gain range to stabilize objective values. Next, we demonstrate performance of our controller. by numerical simulations, feasibility and implementability is confirmed.

2. Circuit model

Let us consider a simple interrupt chaotic system [7] shown in Fig. 1 as an example. The switch is flipped by a certain rule depending on the state and the period. Assume that \( v \) is the state variable, and then the normalized equation is given as follows:

\[
\frac{dv}{dt}(t) = v(t) - E \\
\text{If } t = n\tau \quad \text{then } E = E_{in} \\
\text{If } v(t) > E_r \quad \text{then } E = 0
\]  

(1)

where \( E_{in} \) and \( E_r \) are a direct current bias and a switching threshold value, respectively. \( \tau = T/RC \) is the period of the clock pulse input. If the Poincaré section is defined as \( \Pi = v(t) \in \mathbb{R}; t = n\tau \), trajectories strike two types of solutions, and they can be solved exactly,

Figure 1: DC-DC converter model circuit.
see [3] Therefore the system can be discretized by the Poincaré section, and redefined as follows:

\[ v_{k+1} = g(v_k) = \begin{cases} 
(v_k - E_1) e^{-\tau} + E_1 & v_k < D \\
E_r v_k e^{-\tau} & v_k \geq D 
\end{cases} \quad , \tag{2}
\]

where,

\[ v^* = \frac{-E_r E_1 e^{-\tau}}{E_r (1 - e^{-\tau}) - E_1}, \quad A = \frac{\partial f}{\partial v^*} = \frac{E_r}{E_r - E_1} e^{-\tau}, \quad \text{and} \quad B = \partial f/\partial \lambda. \]

The conditions of stability for the target value \( x^* \) is \(|\mu| < 1\). Therefore the controller gain \( c \) that \( x^* \) becomes stable is derived as follows:

\[ c = \frac{q - A}{B}, \quad -1 < q < 1 \tag{7} \]

A chaotic attractor and UPO with parameters \( E_1 = 3 \) and \( \tau = 0.606 \) are shown in Fig. 2. It is confirmed that the ripple voltages of the periodic orbit are about 0.6V smaller than the chaotic attractor. This model show various responses dependently on the output voltage \( E_r \), and there is the case of large ripple voltages. Unstable periodic orbit similar to the Fig. 2 (b) is embedded into the chaotic attractor. If it can be stabilized by appropriate controlling input, the ripple voltage will be decreased.

![Chaotic attractor](image)

(a) Chaotic attractor \( E_r = 2.5 \)

![Periodic orbit](image)

(b) Periodic orbit: \( E_r = 1.81 \)

Figure 2: Chaotic attractor and periodic orbit with \( E_1 = 3 \) and \( \tau = 0.606 \).

3. Controlling methods

In this paper, two types scheme for stabilization of UPOs are proposed based on a feedback control. The pole assignment method is used for design of a controller gain. Let us consider a 1-dimensional discrete time dynamical system as follows:

\[ x_{k+1} = f(x_k, \lambda), \quad x \in R, \lambda \in R. \tag{4} \]

Assuming the controlling input \( u_k = c(x^* - x_k) \) for the parameter, the system is described as \( x_{k+1} = f(x_k, \lambda + c(x_k - x^*)) \), where \( x^* \) is a target value, and \( c \) is the controller gain. Thus, the characteristic equation is shown as follows:

\[ \chi(\mu) = A + Bc - \mu = 0, \tag{5} \]

with \( v_k < D \). The lower map of Eq. (2) is used for the objective one-periodic orbit obviously. the partial derivative of the map with (2) respect to the period \( \tau \) is given as follows:

\[ B = -E_r \frac{v^* - E_1}{E_r - E_1} e^{-\tau} \tag{8} \]

From Eq. (7) and (6), the ranges of stabilizable gains are determined without experimental results, and they are shown as follows:

\[ -E_r + E_1 + E_r e^{-\tau} < c < \frac{E_r - E_1 + E_r e^{-\tau}}{E_r (v^* - E_1) e^{-\tau}} \tag{9} \]

\[-1.7 < c < -0.79 \tag{10} \]
3.2. Occasional applied pulse input: forcing the interrupt event at appropriate time ahead original clock pulses

If the target circuit is already constructed, and controller can not adjust the frequency of the clock pulse, previous scheme does not apply to one. Next, let us consider the controlling scheme for the circuit with fixed clock pulses. The sketch of the controlling scheme is illustrated as Fig. 4. On the previous controller, it is designed based on changing the circuit phase forcibly to the charge at an appropriate time. On the other hand, this controller changes the mode to the discharge by appropriate pulse. The pulse is added to the RS-FF before the radical clock pulse. Thus, this controller does not influence the clock pulse, and it can apply to the circuit with the fixed clock pulse.

Figure 5 shows behaviour of the circuit with controlling inputs, it can confirm that four type trajectories exist dependently on the initial voltage \( v_k \). (i) and (ii) are the same to the circuit’s one without controller. (iii) and (iv) are new behaviour with influences of applied pulses. They can be derived as two operation mapping with \( \tau - u_k \) and \( u_k \). Therefore, it can be described as \( v_{k+1} = g(g(v_k, \tau - u_k), u_k) \). As a result, the differential equation of the circuit with controller is described as follows:

\[
v_{k+1} = g(g(v_k, \tau - u_k), u_k)
\]

are described as

\[
\begin{align}
(v_k - E_1)e^{-\tau} + E_1 & \quad \text{(i)} \\
E_r \frac{v_k - E_1}{E_r - E_1} e^{-\tau} & \quad \text{(ii)} \\
E_r \frac{v_k - E_1}{E_r - E_1} e^{-\tau} + (1 - e^{-u_k})E_1 & \quad \text{(iii)} \\
\left\{ \frac{E_r}{E_r - E_1} \right\}^2 (v_k - E_1)e^{-\tau} - \frac{E_r}{E_r - E_1} e^{-u_k} & \quad \text{(iv)}
\end{align}
\]

Figure 4: Sketch of Occasional applied pulse events. The red arrow means an additional interrupt pulse.

When \( \xi << 1 \), the input \( u_k \) is also approximately zero. In this case, the map (iii) is applied, and the partial derivative of the map respect to the controlling input \( u_k \) is given as follows:

\[
B = e^{-u_k}E_1 \big|_{u_k=0} = E_1
\]

From Eq. (7) and (6), the ranges of stabilizable gains are determined without experimental results, and they are shown as follows:

\[
-\frac{1}{E_1} - \frac{E_r e^{-\tau}}{E_1(E_r - E_1)} < c < \frac{1}{E_1} - \frac{E_r e^{-\tau}}{E_1(E_r - E_1)}
\]

Note that, if \( u_k < 0 \), the applied pulse have no meaning because the clock pulses arrive before controlling pulse. Therefore, the limitation \( 0 < u_k \) for the controller is necessary, and pulse is applied occasionally to the objective circuit with only \( u_k \) is positive values.

4. Controlling result

Figure 6 shows controlling results. The under arrows are clock pulses applied controlling input values, and the wave form means the capacitor voltage \( v \) of the circuit. Under figures are amounts of controlling inputs. In Fig. 6 (a), it is confirmed that the frequency of clock pulse is decreased at the beginning of the simulation. However, it gets back the ideal value \( \tau \) with decreasing the controlling input value \( u \). Finally, the voltage converged as one-periodic orbit with small ripples. In Fig. 6 (b), the frequency of the clock pulse is not changed, but additional pulse (red arrows) are added, and interrupt events are forced ahead original clock pulses. Under figures show amounts of the controlling inputs. When \( v \) converges to the objective periodic orbit, the input \( u(t) \) also converges to zero. Note that, there is the interval with \( u = 0 \) and no additional pulses, because this scheme has limitation for the amount of controlling input. If the input \( u \) takes negative value, controller applies the additional pulse to the RS-FF after the clock pulse, and it has no meanings. Hence, when the controlling input takes negative value, it becomes zero. Thus, these behaviors are unstable periodic orbits and they are stabilized by controller.
On comparison these result, it seems that the method 2) can control the circuit to the objective UPO from simulation results. Actually, setting times of two methods are $t_{s1} \approx 15$ and $t_{s2} \approx 4.5$, respectively. However, on the view point of robustness, the method 2) has limitation for the amount of controlling input, and it causes the degration of robustness. The method 1) can applyes bipolar inputs, and can respond unexpected motions (offcourse, this method has also limitation for the controlling input, and it is $u_k < \tau$, but it is slacker than the method 2)). Thus, the method 1) has high robustness than the method 2). Note that, these characteristics is changed by adjusting the parameter $q$ of the controller, and this paper does not discuss appropriate vales of them.

5. Conclusion

In this study, we propose two types of chaos controlling for hybrid dynamical systems, and demonstrate the performance of them. As a result, our controller can stabilize unstable periodic orbits or DC-DC converter model circuits, and can reduce the ripples of them.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 16K18107.

References


Circuit Characteristics of DC/DC Converter for Waste Heat Recovery Power System

Kouji Nakasuka†, Nobuyuki Kasa†, Hiroyuki Asahara†, Kenichi Takarabe†, Masahiro Minowa‡, Tsutomu Iida*

†Okayama University of Science
1–1 Ridai-cho Kita-ku, Okayama-shi, Okayama 700-0005, Japan
‡SWCC Showa Cable Systems CO.,LTD.
4-3-1 Toranomon Minato-ku, Tokyo 105-6012, Japan
*Tokyo University of Science
6-3-1 Shinjyuku Katusuka-ku, Tokyo 125-8585, Japan

Email: t16em03nk@ous.co.jp, kasa@ee.ous.ac.jp, asahara@ee.ous.ac.jp,
takarabe@das.ous.ac.jp, m.minowa243@cs.swcc.co.jp, iida_lab@mac.com

Abstract—In this report, the multi input flyback converter for waste heat generation systems is proposed. This research is aiming for maximizing the output power and high efficiency of converter circuit design, high efficiency of power converting circuit and the waste heat power generation system.

When the switching frequency is changed, the efficiency of the converter is measured. The experimental results show that the efficiency depends on the switching frequency.

1. Introduction

Thermoelectric devices can convert some of this waste heat into useful electrical energy. Thermoelectric generation is a process where power is obtained by giving a temperature difference to the thermoelectric module in accordance to the Seebeck effect theory. Waste heat power generation system is a generator for recovering electrical energy by giving a temperature difference in the thermoelectric module (TEM), but the energy conversion efficiency and durability problems are often in the spread. Therefore, this research is considering the application of a thermoelectric device in extent to a variety of waste heat and study about the multi-input flyback converter for waste heat power generation system.

This research is aiming for high efficiency of converter circuit design, high efficiency of power converting circuit for waste heat power generation system.

2. Multi Input Flyback Converter

Power converter as a means of power conversion from source to load need to be properly controlled and regulated in order to achieve stability at all operating conditions and environments. Experiment is been carried out based on the basic configuration of flyback converter. Figure 1 shows conventional converter. The voltage across each TEM is very low (0.3V). To obtain the higher voltage for the system, the modules must be series-connected. To implement the system, it requires a DC/DC converter. TEMs are series-connected. The current is fixed by the module which is located on low temperature.
Figure 2 shows our proposed multi input flyback converter used for this research. The modules are parallel-connected. The current is not fixed by the module which is located on low temperature.

Figure 3 shows the comparison efficiency graph of when the input voltage of the dc voltage power supply is changing from 0.4V to 1.0V. When the switching devices are changed from IRF520N to IRL2505 whose on-resistance is very low, the efficiency is improved at 3 point.

3. Experimental System

An experimental system was assembled in the laboratory as shown in Figure 4. The converter consists of the main switches (MOSFETs) and the flyback transformer. The system is implemented by a micro controller.

Figure 4: Experimental system.

<table>
<thead>
<tr>
<th>TEM #1</th>
<th>TEM #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voltage</td>
<td>0.20V</td>
</tr>
<tr>
<td>Current</td>
<td>0.26A</td>
</tr>
<tr>
<td>Power</td>
<td>0.05W</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Electric load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voltage</td>
</tr>
<tr>
<td>Current</td>
</tr>
<tr>
<td>Output Power</td>
</tr>
</tbody>
</table>

Figure 5 shows the experimental results of the system at 300 degrees centigrade. The efficiency is 66%.

Figure 6 shows the relationship between the temperature of the hot side of the thermoelectric module and the power which is supplied by 12V power supply.

4. Conclusions

The prototype of waste heat recovery power generation system based on multi-input flyback converter has been proposed.

5. References

Research Related to Power Packet and Its Application

Takashi Hikihara

†Electrical and Electronic Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615-8510, Japan
Email: hikihara.takashi.2n@kyoto-u.ac.jp

Abstract—In the Cross-ministerial Strategic Innovation Promotion Program (SIP), a project named “Next Generation Power Electronics” is ongoing. In the project, our group is figuring out the research towards power processing through the high frequency power switching and its applications. In this report, the outline and the target of the project is briefly explained.

1. Introduction

One of the research topics of the Cross-ministerial Strategic Innovation Promotion Program (SIP) is the project “Next Generation Power Electronics.” Since 2014, 10 topics are on going in SIP. We have engaged in the feasibility study of “Research and development of low power and small size integrated power circuit and power processing technology based on applications of SiC power devices.” This special session exhibits some results of the project and encourage researchers to approach non-conventional research filed with nonlinear thinking.

The project advances the applications of wide bandage semiconductors, SiC and GaN based on their physical superior characteristics. In particular, power electronics technology at the high frequency switching are not matured more than 10 MHz. The power switching in the range has not been well developed until the appearance of SiC and GaN power devices. In the range the power switching also shows the ability of the transfer of information.

Our group have proposed the new method of power transfer with information technology in the physical layer. Power is packetized as an unit and discretized power is transfer with each information tag. We call the technology of “power processing.” In the following, we explains the concept of the power processing.

2. High Frequency Switching and Power Packet

Power electronics is the technology of converting waveform, quantity, or speed (frequency) on loads through the circuit switching. Then the voltage/current on loads are switched, so that the averaged voltage/current will be regulated according to the duty of switching rate. The ratings of the power devices govern the ability of the conversion. At the low frequency, inductors and capacitors, which are called storage passive devices, are inevitable for keeping continuous current and voltage. At the low switching frequency they becomes huge and they decides the size of the convertor circuits. From the view point of high power conversion, the passive devices also becomes huge.

Apparently, the high frequency switching makes it possible to deduce the size of the passive devices, because the power stored becomes low due to the intermittent duration. On the other hand, the switches show the physical limitation on the power rating, temperature limitation, and switching speed.

Recent development of wide bandage semiconductors changes the state of power switch. The show the possibility of power switching at the range of radio frequency, that is more than 10 MHz. They are enough to send information. At the sometime, the switch devices possess a capability of high current and the tolerance to high voltage with low on resistance.

Here we are proposing a system which transfer electric power at a pulse waveforms with an information-tags attached to the power. Needless to say, it becomes possible because of the appearance of the wide bandage semiconductors and their power devices.

3. Power Packetization and Processing

When electric power is packetized, total power is given by the accumulation of unit pulse power. The density of the packet changes the power density on time. This is a type of pulse density modulation (PDM). The method is completely different from PAM and PWM, which are analog conversion by switching. Their conversion depend on the duty in a time duration. However, PDM does not depend on the duty of pulse width, but the density of single pulse in a time duration. The density decides the power of the conversion.

![Power packet structure](image-url)
As mentioned above, the power packet is accompanied by the tag. The information is tightly attached to physical power for each power unit \[1\]. The destination of their pulse power are decided according to the tag and the node devices, which are “mixer” and ’router’, are operated for sending and receiving. The synchronization of the router is also achieved by the preamble of tag. The network consisted of the routers are physical network of power transfer without tight physical connections between power sources and loads. Fig. 2 is an example of the power packet dispatching network \[2\].

.png

Figure 2: Power packet dispatching network

We have already discussed the system based on the communication theory \( \text{Y} \) citeshannon and the differences are clearly shown \[4\].

4. Formulation of Power Packet Dispatching

The power packet is defined as a wave of power with information. As for the physical layer, various configurations are possible with keeping this concept. However, the substantial characteristic is that the power is quantized and transferred by the density distribution in a duration of time between two nodes. That is, power is digitized and quantized. In the quantized form, the spatially transferred power is decided between nodes \[2\]. In and out flow of power at a node corresponds to a density of packet. The deference of the densities implies the stored power at the node and loss. As for the link, the power is kept between nodes. We would like to represent the system based on the wave equation and quantum mechanics.

Here, we define the density (probability) of power packet by \( \sigma_n(t) \) at a node \( n \) in the \( s \)-th time duration \[ t_s, t_s + \Delta t_s \], \( v_n(t) \) is set as the voltage of packet, and \( i_n(t) \) the current through the node by the packet. The energy \( u_n(s) \) is given in the duration by the following relation:

\[
u_n(s) = \int_{t_s}^{t_s + \Delta t_s} \sigma_n(t) \cdot v_n(t) \cdot i_n(t) \, dt. \tag{1}\]

Hereafter, the system is normalized by an appropriate base.

The total flow of power from the \( k \)-th node depends on the total flow into the node. The spatial node index is given by subscript. If there is no storage (buffer) at the node, the energy conservation law at \( t_s \) is represented by

\[
\sum_{i=[r-k]}^{r} u_i^k(s) = \sum_{j=[r-1]}^{r} u_j^k(s). \tag{2}\]

The density of packet at \( k \) in a time duration implies the energy flow through \( k \) as follows:

\[
u_k(s) = -\int_{t_s}^{t_s + \Delta t_s} \sigma_k(t) \cdot v_k(t) \cdot i_k(t) \, dt. \tag{3}\]

Here, assume the single connection of nodes, then the energy flow by packets is defined between node \( k \) and \( (k + 1) \) as

\[
u_k^{k+1}(s) = -u_k^{k+1}(s) + u_k(s) \tag{4}\]

In the connection, \( u_k^{k+1}(s) = 0 \) implies lossless line. Between nodes, the relationship holds:

\[
u_k^{k+1}(s) = -\int_{t_s}^{t_s + \Delta t_s} (\sigma_{k+1}(t) \cdot v_{k+1}(t) \cdot i_{k+1}(t) - \sigma_k(t) \cdot v_k(t) \cdot i_k(t)) \, dt. \tag{5}\]

The power packet is quantized in voltage. At the setting, \( v_k+1(s) = v_k(s) \). When \( \sigma_{k+1}(s) = \sigma_k(s) \), the transferred energy from node \( k \) to \( k + 1 \) at \( t_s \) becomes

\[
u_k^{k+1}(s) = -v_k(s) \int_{t_s}^{t_s + \Delta t_s} \sigma_k(t) \cdot (i_{k+1}(t) - i_k(t)) \, dt. \tag{6}\]
The relationship implies that the energy between nodes depends on the current through the neighboring loop.

\[
\sigma_k = \int_{t_k}^{t_{k+\Delta t}} \sigma_k(t) \, dt
\]  

(7)
gives the temporal rate of packet.

The total energy of the system is given by \( U(s) \). It normalizes the system as

\[
\dot{i}_{k+1}(s) = -\frac{\sigma_k^{k+1}(s)}{U(s)} = \frac{A \sigma_k}{K} v_k(s) \cdot [i_{k+1}(s) - i_k(s)]
\]  

(8)

Here \( A \) is the normalizing coefficient and \( K \) the coupling inductance. The \( \Gamma \)-shaped radder connection of \( L \) and \( C \) is the unit element of distribution line. Then,

\[
i_{k+1}(s) - i_k(s) = v_k(s) \cdot i_0 C.
\]

Here the current can be replaced by voltages of nodes. Real value \( v_k \) can be replaced by complex value with the complex conjugate voltage \( v_k^* \). And the current

\[
\dot{i}_{k+1}(s) \sim \frac{i_0 C A \sigma_k}{K} (v_k^*(s) + v_k^*(s)) [v_k(s) - v_{k+1}(s)]
\]

\[
- (v_k^*(s) - v_{k+1}(s))]
\]  

(9)

5. Remarks

This paper summarizes the engineering and mathematical basis of the research supported by SIP project. The idea of power processing has already been proposed in several research project, including NICT project and Super Cluster Program, with achievement of power packetization in the physical layer. The conventional power transfer completely included in the formulation of power packet dispatching. The each power sources faces to another types of dynamics with combination to ICT.

In the packetized power transfer, the optimization of the packet distribution and flow control will be an important target in the next step. The realization of system completely depends on the development of power devices and their circuit implementations at high frequency switching range.

Acknowledgement

The research results were based on the results of the project supported by NICT. It is partially supported by SIP project funded by NEDO.

References


A Packet-based Energy Management System for Multi-legged Walking Robots

Yusuke Kojima\textsuperscript{\textdagger}, Shun-ichi Azuma\textsuperscript{\textdagger}, Taichi Kitao\textsuperscript{\textdagger}, Ichiro Maruta\textsuperscript{\textdagger}, and Toshiharu Sugie\textsuperscript{\textdagger}

\textsuperscript{\textdagger}Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
Email: kojima@robot.kuass.kyoto-u.ac.jp, sazuma@i.kyoto-u.ac.jp, kitao@robot.kuass.kyoto-u.ac.jp, maruta@i.kyoto-u.ac.jp, sugie@i.kyoto-u.ac.jp

Abstract—Recent years, packet-based energy management systems attract attention as a new technology. In order to demonstrate such systems in the real world, we are interested in applying a packet-based energy management system to a multi-legged walking robot actuated by multiple motors. This paper reports our results on the hardware implementation and an experimental result.

1. Introduction

Power packets are formatted units of electricity, containing power and information on delivery. It is known to be a new concept of electric power supply and is expected as a future technology.

For power packets and systems implementing them, several results have been obtained so far (see [1] and references therein). On the other hand, although the power packets are expected to be applied to autonomous robots, there is no result on hardware development and experiment.

In this paper, we report on hardware development of a packet-based energy management system for an autonomous multi-legged walking robot and an experimental result. These results demonstrate the feasibility of the concept of power packets and the effectiveness in the real world.

2. Concept of a Packet-based Energy Management System

The concept of our packet-based energy management system is shown in Fig. 1. The aim of this system is to supply a power packet, i.e., intermittent power, to \( n \) loads for achieving a desired task in the whole system. In this system, a power packet can be sent to a single load in each time step.

The power dispatching procedure is as follows. First, each packet receiver estimates the required amount of power for the corresponding load and sends a request to the power router. Next, the power router determines the destination of the power packet and sends a request to the power. Then the power sends the energy to the power router, and the router supplies it to the destination.


We have developed a packet-based energy management system for a multi-legged walking robot [2] (Kondo Kagaku Co., Ltd.). This robot has 6 legs and each leg equips 2 motors. These motors are servo motors whose max speed is 0.14s/60°. In order to achieve walking, we need to send power packets to the 12 motors in an appropriate manner on the basis of the predictions of power demand. In the following subsections, we explain the packet receivers and the power router for the walking robot.
3.1. Packet Receiver

The packet receiver is shown in Fig. 3. This is mainly composed of a capacitor and a diode. This receiver generates a request signal based on the terminal voltage of the capacitor. If a power packet is provided from the power router, this receiver saves power packet in the capacitor and supplies power to the corresponding motor.

3.2. Power Router

The power router is implemented by Arduino Mega 2560 [3] and Photo MOS Switch [4], as shown in Fig. 4. The router receives a request signal from the packet receivers, and selects a receiver which needs power the most. Next, the router actually sends a power packet to the destination.

4. Experimental Result

We have confirmed that the robot walks at the speed of approximately 0.04 m/s. Fig. 5 shows the snapshots of the walking.

5. Conclusion

We have developed a packet-based energy management system for a multi-legged walking robot, and demonstrated its performance. One of the important future topics is to develop a more efficient method to dispatch power packets.

Acknowledgments

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program from New Energy and Industrial Technology Development Organization.

References

Design and Implementation of a Routing Protocol for Power Packet Network

Yusuke IKUMA†, Thong HUYNH†, Kaori KURODA† and Mikio HASEGAWA†

†Department of Electrical Engineering, Tokyo University of Science
6-3-1 Nijuku, Katsushika, Tokyo, 125-8585, Japan
Email: [ikuma, huythongtc]@haselab.ee.kagu.tus.ac.jp, {kuroda, hasegawa}@ee.kagu.tus.ac.jp

Abstract—In this paper, we study the routing protocol for power packet networks. In the conventional routing protocol, since the router needs to read all header information before forwarding packet to other routers, the network might suffer high latency when the networks size is increased. We proposed the new protocol to reduce the processing time for routing at the router. In our proposed protocol, the router does not need to read entire header information for forwarding packet. Therefore, the processing time of the router is reduced. We implement the proposed protocol on a micro-computer board, which can be used as a controller of the power packet router. The simulation results show that our proposed protocol can not only properly operate in power packet networks but also provide lower latency than the conventional routing protocol.

1. Introduction
In recent years, as a new application area of the wide band gap semiconductor, small integrated circuit that is switchable on high voltage and high frequency is under developing by using power device like SiC or GaN. It is expected to transmit energy in packets. The power packet network that enables energy transfer [1] by the power packets has been developed in Ref. [2]. The sources transfer power packets to arbitrary load via power packet routers.

In this paper, we design the routing protocol for power packet network. Further, we implement the proposed protocol on a micro-computer board as a controller of the power packet router, and verify its operation. In conventional protocol, each router needs to read all header information before forwarding packet. This might lead to the increasing of the delay when the networks size is increasing. Furthermore, it is difficult to directly apply routing protocols in communication networks for transmitting power packet, because the bit rate on the power packet network is significantly slow compared to the communication networks. Therefore, it is necessary to develop low-latency protocol at a low bitrate.

We design the protocol for power packet routing with low delay. Our protocol is based on the protocol in the layer 2 of the OSI reference model (data link layer). To validate the operation of the networks, we implement the proposed protocol on a micro-computer board as a controller of the power packet router.

2. Design of the protocol in power packet routing
In this paper, we design the protocol for the power packet router developed in [2]. The power packet delivers electric power according to the attached information (such as the destination address) superimposed on the voltage waveform. A schematic diagram of a power packet router developed in the literature [2] is shown in Fig. 1. When the power packet comes to the router, the header information will pass through isolator to the controller. Based on the destination address in the header, the controller orders the gate driver to forward power packet to the destination. The gate driver controls the switch circuit to send the power packet via output port. Finally, the power packet router attaches the header information based on the original one stored in the memory.

Fig 1: Schematic diagram of power packet router. [2]

In the routing algorithm in the literature [2], after reading the entire header information, the router checks the destination address in the header and their own routing table. Then, the router determines the destination port and forwards a packet. For example, in Fig.2, after reading all information about the destination address (DA) and source address (SA) and checking with routing table, the router will forward packet to destination router via corresponding port. However, as the increasing of network size, the header length becomes longer and the router takes more time to read the information. The power packet network might suffer high latency.

Therefore, we design the routing protocol referred to TRILL (Transparent Interconnection of Lots of Links)
standardized in RFC6325 [3] and RFC6326 [4] by IETF, to provide a low-latency routing. This routing protocol operates at layer 2 of the OSI reference model (data link layer), and forwards the packet efficiently by implementing in the inter-router communication.

As shown in Fig. 3, we add an identifier that includes destination router identification (DR) and source router identification (SR) for the router before the destination and source address information. The number of identifier for the router is much shorter than the number of individual addresses assigned to all routers in the network. Therefore, in the inter-router communication, the router starts to forward packet immediately after reading the router identifier of the header and checking the DR and the routing table. In our proposed protocol, because the router does not need to read entire header information for forwarding packet, the processing time at the controller in the router is small. Therefore, our protocol protocol can reduce the latency of the networks, especially when the number of router in network is large.

3. Implementation of the protocol for power packet routing

We implement the proposed protocol in an open source hardware Arduino, which can be used as a controller of the power packet router. We validate the proposed protocol and compare the performance with a simple conventional protocol.

3.1. Prototype format of the protocol for power packet routing

In carrying out the implementation of the protocol for power packet routing, we propose a prototype format of the power packet referred to the OSI reference model and TRILL in Fig 4.

First, we add the preamble 12 bits for synchronization to the top. The following 8 bits are dedicated for the source and destination router ID (each 4 bits), which cover for 16 routers. The following 3 bits are indicated the QoS (Quality of Service) that corresponds to ToS (Type of Service) in IP. These bits are used for implementing priority control and bandwidth control. The following 16 bits are the destination address and the source address (each 8 bits), which cover 256 loads. Finally, three bits are used for Type to indicate the protocols such as IPv4 or IPv6 in the upper layer.

3.2. Network configuration in the experimental environment

We setup the experimental environment as in Fig. 5. Our power packet network includes one source, 4 routers to forward the packet. We use three different LED lights as the load in the experiment. The bit rate is set to 20 Hz, the packet transmission interval is set to 3 seconds, and the packet format is as shown in Fig. 4.

As operation check, we assign different addresses to red, yellow and blue LED lights, and transmit the power packet from source. We measure the voltage at the router with oscilloscope to confirm that the router starts to transmit the power packet immediately after reading the DR in the header.
3.3. Experiments on proposed routing protocol

3.3.1. Waveform at transmitting the packet to red LED

Fig. 6 shows the waveform of the transmitting packet from RT-A to red LED. The points 1, 2, 3 in Fig. 6 are the observation points. The waveforms of observation points 1 and 2 show that the packet is started to transmit immediately after reading the destination router ID.

3.3.2. Waveform at transmitting the packet to yellow LED

Fig. 7 shows the waveform of the transmitting packet from RT-A to yellow LED. The waveform of observation points 1 and 2 show that the packet is started to transmit immediately after reading the destination router ID.

3.3.3. Waveform at transmitting the packet to blue LED

Fig. 8 shows the waveform of the transmitting packet from RT-A, RT-C, RT-D to red LED. The points 1 to 4 in Fig. 8 are the observation points of the power packet. Although in this route, we transmit the packet over longer path than the previous, the packet is also started to transmit immediately after reading the destination router ID.

3.4. Comparison of the waveforms in the conventional method and the proposed method

We also implement the conventional method on Arduino for comparison. In the conventional method, we use the packet format which is removed the destination router ID and the source router ID from the packet format of Fig. 4. On this setting, we observe the waveform at transmitting the packet to blue LED, and compare two methods. Fig. 9 shows the waveform of the conventional method and the proposed method. And the waveform of the proposed method is same as Fig 8.

From the comparison of the two waveforms shown as in Fig. 9, the proposed header is 8 bits longer than the
conventional method, however the time until the power reaches the red LED is shorter than the conventional one. That is because the processing time of the router in our proposed protocol is reduced.

4. Conclusion

In this paper, we designed and implemented a routing protocol for the power packet networks. In order to reduce the processing time at each router, we designed a routing protocol based on TRILL. Since each router starts to transmit power packet immediately after reading the DR in the header, our proposed protocol provides low latency for the networks.

By implementing the proposed protocol, we confirmed that our proposed protocol can operate properly in current designed power packet networks. Furthermore, we also achieve lower latency than the conventional method.

In this paper, we have implemented the controller part for verifying the operation of the proposed protocol. Currently, we are trying to implement the proposed protocol on a power packet router system.

This work was partially supported by the Cross-ministerial Strategic Innovation Promotion Program from New Energy and Industrial Technology Development Organization, Japan.

References


A simulation study of an algorithm for distributing power packets in a network

Hiroyasu Ando, Shun-ichi Azuma and Ryo Takahashi

†Division of Policy and Planning Sciences, University of Tsukuba, Ibaraki, Japan
‡Graduate School of Informatics, Kyoto University, Kyoto, Japan
§Graduate School of Engineering, Kyoto University, Kyoto, Japan
Email: ando@sk.tsukuba.ac.jp, sazuma@i.kyoto-u.ac.jp, takahashi.ryo.2n@kyoto-u.ac.jp

Abstract—Power packet dispatching system is an efficient technology for energy saving. To support this system, we propose an algorithm for distributing power packets in a network system. In this study, we consider the consensus dynamics in a network by assuming the state variable at each node as an amount of power. In addition, we assume that network topology changes in time. These assumptions are effective for distributing power packets in power network systems in terms of decentralized control.

1. Introduction

Power packet dispatching systems have been one of promising technologies in the field of power electronics for saving power consumption [1]. Power packets are discrete units of voltage and transmitted with header and footer tags in which an information of the destination node in a power network is retained as the internet packets do. The transmitted power packets are delivered to a destination node by a router assigned at each node. In the power packets routing, there exists several limitations intrinsic to electronic circuits and devices. Therefore, a mathematical study with models is useful so as to consider an efficient algorithm for distributing power packets in a network. Specifically, we focus on consensus dynamics [2] as a simple model of diffusion dynamics in a network, which corresponds to delivering process of power packets in a power network. Then, we propose a simple algorithm for distributing power packets in the network to satisfy target amount of power in each node. That is to say, the amount of power at each node is controlled to a different level of power from the equal amount. Moreover, to diversify distribution patterns of power packets, we consider switching networks in which accessible links change with time [2]. In particular, we focus on the condition that a node in a network can interact with only one randomly selected node in the network.

2. Numerical simulations

Figure 1 shows time series of consensus dynamics at all the nodes in a 4 node network controlled by the algorithm. As shown in the time series, the state of each node is converging to a certain value, which implies that all amount of power is distributed to every node towards targeting level of power. In fact, the error from the controlled state is less than 10% in Figure 1.

3. Summary

We have numerically investigated the consensus dynamics with switching topologies of networks controlled by a local demand of power at each node. Our main concern of this study is how a decentralized control can be combined with central one for saving energy and time consumption. This combination should be done in a future work especially as circuits implementation of the proposed algorithm.

Acknowledgments

The authors would like to thank CSTI, SIP, “Next-generation power electronics” (NEDO).

References

A study of pulse shape suitable for power packet dispatching on power transmission line

Ryo Takahashi*, Takuya Kajiyama†, Takafumi Okuda‡, and Takashi Hikihara§

*†§ Department of Electrical Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615-8510, Japan
‡ Department of Electronic Science and Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615-8510, Japan
Email: *takahashi.ryo.2n@kyoto-u.ac.jp

Abstract—Power packet dispatching system is proposed as one of the advanced power distribution systems, which is suitable for smart energy management. In the system, electric power is delivered by a packetized voltage formed as pulses. In this paper, the pulse shape suitable for power packet delivery is investigated. Here, two types of pulse waveforms are adopted, i.e., rectangular and gaussian. In the paper, we investigate the pulse distortion and propagation loss through power transmission line for these shapes. As results, it is found that the gaussian pulse can suppress more than the rectangular one.

1. Introduction

Recently, advance power distribution systems are proposed, which utilize gathered information for more effective power management. Highly developed Information and Communications Technologies (ICTs) are essential to realize these system, which realize the large information gathering, networked control, pattern analysis for consumers, etc. In addition, new wide-band gap power devices such as SiC and GaN which can manage high-power and reduce power loss starts to be produced. These power devices have a potential to realize significantly lower switching power loss, higher frequency switching, and higher power capacity with compared to Si devices. It is also undoubtedly essential to adopt these devices and to develop technologies for exhibiting their performance sufficiently in power management.

As one of the advanced power distribution systems, a power packet dispatching system has been proposed [1]. The concept was proposed in 1998 [2]. Recently developed ICT and wide-band gap power devices allow us to extend the concept and development for more practical use. In the power packet dispatching system, electric power is delivered as a pulse with information bits, i.e., a packetized form. The proposed power packet consists of a header, a payload, and a footer. Figure 1 shows the configuration of the power packet. The payload carries the power, which means electric current exists in the duration of payload. As an information tag, the header and footer are attached to the payload as the voltage waveform physically. The header includes the start signal, the address of source and load. The footer includes the mark to notify the end of packet.

2. Evaluation of pulse distortion and propagation loss on a power line

In this paper, the pulse shape suitable for power delivery as a payload in a power packet is investigated. The propagation characteristics of power transmission line depends on the frequency [3, 4]. The power packet dispatching is proposed with a band of several hundred kHz to several MHz. Obviously, it is desired to suppress power loss in the power line. The pulse distortion should be also suppressed to prevent the damage to the apparatuses. The spectrum should be limited to prevent the interference. Therefore, the suitable pulse shape for power delivery should be selected. Here, we focus on the rectangular and Gaussian pulse shapes. The pulse shape in the prototype of power packet dispatching system is rectangular [1]. It is because the waveform of power packet has been obtained by the ON/OFF switching of connected dc power source. As is well known, the rectangular shape includes higher harmonic spectrum. On the other hand, the spectrum of Gaussian pulse is limited. Here, the pulse distortion and propagation loss for these pulse shapes are investigated numerically and experimentally.

Figure 2 shows the schematic diagram of the system investigated here. This system consists of a pulse power source, a load, and an electrical cable with two wires. The
quantities $Z_s$, $Z_0$ and $Z_L$ indicate the internal impedance of the source, the characteristic impedance of the line, and the impedance of load, respectively. As a power line, we adopt a 20 m VVF (Vinyl insulation, Vinyl sheath, Flat) electrical cable which is widely used for indoor power lines in Japan. Here, the channel model of this cable proposed in [3] is adopted, and the resistive load is used. We set $Z_s = Z_L = 50$ $\Omega$. The quantity $Z_0$ is not the same with $Z_s$ or $Z_L$, so that this system contains the impedance mismatching which causes the reflection of voltage or current.

Here, the switching frequency is set at 1 MHz, duty ratio $D = 0.5$. Figure 3 shows the waveforms at source and load for each pulse shape obtained by numerical simulations. The results show that Gaussian pulse can suppress the pulse distortion rather than rectangular one.

The efficiency of power delivery via a power line for each pulse shape is also evaluated. High efficiency implies low power loss. Note that, in this study, the switching loss is ignored. Figure 4 shows the efficiencies obtained by numerical simulations. In the figure, the one for sinusoidal wave, i.e. ac power delivery, is also depicted. From the results, it is found that the appropriate shape depends on frequency. Lower than 1 MHz, from the efficiency point of view, the Gaussian pulse is a bit better or almost the same with the rectangular pulse.

3. Conclusion and discussion

In this study, a suitable pulse shape for power delivery in power packet dispatching system was investigated. We evaluated the pulse distortion and power loss for the rectangular and Gaussian pulse shapes. The results show that the Gaussian shape can suppress the pulse distortion. As for the power loss in the power line, lower than 1 MHz, the Gaussian shape is a bit better than the rectangular one. From the viewpoint of interference noise, the Gaussian shape is better because its spectrum is limited.

Acknowledgments

This work was partially supported by Cross-ministerial Strategic Innovatoin Program from New Energy and Industrial Technology Development Organization, Japan, and by the Super Cluster Program from Japan Science and Technology Agency. R. Takahashi was partially supported by the Japan Society for the Promotion of Science Grand-in-Aid for Young Scientist (B) under Grant No. 26820144.

References


Power Conversion Efficiency of RF-DC Rectifier for RF Energy Harvesting

Chang-Jun Ahn†, Takeshi Kamio‡, Hisato Fujisaka‡, and Kazuhisa Haeiwa†

†Graduate School of Engineering, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522 Japan
‡Faculty of Information Sciences, Hiroshima City University
3-4-1 Ozuka-higashi, Asaminami, Hiroshima, 731-3194 Japan
Email: junny@m.ieice.org

Abstract—RF energy harvesting technology is strongly depended on the RF-DC rectifier circuit in the wireless power transmission. In this paper, we show the power conversion efficiency of developed RF-DC rectifier circuit for RF energy harvesting.

1. Introduction

Recently, the availability of free RF energy has increased due to advent of wireless communications and broadcasting systems. RF energy harvesting is the process of extracting small amounts of energy from the ambient environment. This energy can be used to power either portable electronic devices, such as wireless sensing nodes, mobile phones and medical devices, or to charge electrical storage devices (rechargeable battery or capacitor), which can be used at different time intervals for power applications. RF energy harvesting technology is strongly depended on the RF-DC rectifier circuit in the wireless power transmission. The RF-DC rectifier for converting microwave power to DC power has attracted considerable attention in the development of the wireless power transmission [1]. The application of this technology can be used in low power mobile devices, such as radio-frequency identification (RFID) and Zigbee. In this paper, we show the power conversion efficiency of developed RF-DC rectifier circuit for RF energy harvesting.

2. RF-DC rectifier

Figure 1 shows the manufactured RF-DC rectifier circuit. The rectifier circuit is a key element to improve the RF-DC conversion efficiency. A Schottky diode HSMS-2862 was chosen for the rectifying circuit. The inductor and microstrip line are used for input impedance matching circuit and the capacitors are used for the DC block capacitor and the DC filter capacitor. A 2.13 GHz RF-DC rectifier was developed using the previous developed rectenna of 5.8 GHz because the frequencies of 2.12 GHz and 2.14 GHz are the CDMA and WCDMA frequency bands, respectively.

3. Experimental Results

We setup the experiment of power transmission through wires to validate the performance of manufactured RF-DC rectifier circuit. A signal is generated by SMJ100A vector signal generator (ROHDE-SCHWARZ). In this test, the RF power is generated by 8dBm and 11dBm. The RF-DC conversion efficiency is calculated by

$$\eta = \frac{V_{out}/R_L}{P_{RF}}$$

where $V_{out}$ is the measured output DC voltage on the load impedance and $P_{RF}$ is the RF power. Fig. 2 shows RF-DC conversion efficiency versus various input frequency. The vertical axis represents the conversion efficiency. The incident power was varied from 8dBm to 11dBm, and input frequency was varied from 2.1GHz to 2.171GHz. Note that the weaker incident power shows a poor conversion property. The conversion efficiency was improved significantly with increasing incident power. This means that higher incident power can increase the conversion efficiency. The conversion efficiency shows the best performance from 2.13GHz to 2.15GHz due to the load impedance as
Due to the experimental results, we should consider the optimum load impedance with higher incident power case to maximize the RF-DC conversion efficiency.

**4. Conclusion**

From the evaluated results, it is found that the weaker the input power shows a low efficiency of the rectenna. Therefore, it is necessary to increase the input power for efficiency of the rectenna. Moreover, when we adjust the optimum load impedance with higher incident power case, it is possible to maximize the RF-DC conversion efficiency.

**Acknowledgement**

This work was partly supported by the Grant of Scientific Research (C) No. 26420339 from JSPS.

**References**


Performance Enhancement using Unitary Frequency Modulation with PLC and VLC Cooperation

Chang-Jun Ahn† and Yeon-Ho Chung‡

†Graduate School of Engineering, Chiba University
1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba, 263-8522 Japan
‡Department of Information and Communications Engineering, Pukyong National University
599-1, Daeyon3-Dong, Nam-Gu, Busan, 608-707 Korea
Email: junny@m.ieice.org

Abstract—Visible light communication (VLC) is expected as a means for ubiquitous communication using increasingly popular LED lights. However, for indoor optical channel, the received signals are with line of sight (LOS) component but also delayed components due to reflected secondary paths. Therefore, the energy is spread over the time, resulting in a reduction of the converted current at the photodiode. From this reason, the diversity gain of VLC is rarely expected as compared to radio frequency systems. To improve the VLC performance, we consider the combination of power line communication (PLC) and VLC systems for achieving the frequency diversity. PLC systems have many advantages and are assumed to be one of prospective solutions for short distance or in-home communication networks [3],[4]. PLC takes the advantage of use in everyplace at home without additional network line. VLC is easily implemented by using PLC system. Moreover, many electrical appliances frequently cause man-made electromagnetic noise on power line channels. Due to these reasons, the diversity gain of VLC is rarely expected as compared to radio frequency (RF) systems [2]. To improve the VLC performance, we consider the combination of power line communication (PLC) and VLC systems for achieving the frequency diversity. PLC systems have many advantages and are assumed to be one of prospective solutions for short distance or in-home communication networks [3],[4].

1. Introduction

Using increasingly popular LED lights, the visible light communications (VLC) is expected as a means for ubiquitous communication [1]. VLC has some of the noteworthy advantages over radio frequency and infrared systems. In general, visible light does not damage to human body and eyes and VLC has high security of data because the optical signal does not pass the wall, resulting that the communication space is limited to a certain range. Moreover, since LED lighting has recently become part of a building infrastructure, making visible light communication infrastructure fairly easy by adding communication function to LED lighting. As these features, VLC is widely considered as a new way of wireless communication technology.

However, for indoor optical channel, the received signals are with line of sight (LOS) component but also delayed components due to reflected secondary paths. Therefore, the energy is spread over the time, resulting in a reduction of the converted current at the photodiode. Furthermore, the reflected signal power is significantly reduced due to surface reflectivity. From these reasons, the diversity gain of VLC is rarely expected as compared to radio frequency (RF) systems [2]. To improve the VLC performance, we consider the combination of power line communication (PLC) and VLC systems for achieving the frequency diversity. PLC systems have many advantages and are assumed to be one of prospective solutions for short distance or in-home communication networks [3],[4]. PLC takes the advantage of use in everyplace at home without additional network line. VLC is easily implemented by using PLC system. However, power line channel is the time-and-frequency variant and exhibits remarkable difference between locations, according to its network topology, the types of wire lines. Moreover, many electrical appliances frequently cause man-made electromagnetic noise on power line channels. Due to the above-mentioned PLC properties, the performance of PLC system would be degraded. However, using time-and-frequency variance of PLC channel, the diversity gain is also expected. To overcome the above-mentioned problems and improve the system performance of PLC-VLC/OFDM, in this paper, we consider the unitary signal constellation scheme [5],[6].

2. System model

2.1. Indoor Optical Wireless Channel

An electrical current \( y(t) \) at the receiver due to the transmitted optical intensity waveform \( x_{o}(t) \) can be expressed as

\[
y(t) = r\zeta h_{o}(t)x_{o}(t) + w(t),
\]

(1)

where \( r \) is the responsivity of the photodiode at the receiver, \( \zeta \) is the effective receiver area, \( h_{o}(t) \) is the optical channel impulse response, and \( w(t) \) is the noise component. For the free-space indoor optical channel, the received signals are with LOS component but also delayed components due to reflected secondary paths. Therefore, the energy is spread over the time, resulting in a reduction of the converted current at the photodiode. Furthermore, the reflected signal power is significantly reduced due to surface reflectivity. From these
reasons, the diversity gain is rarely expected as compared to RF systems. Therefore, the combination of PLC and VLC is considerable approach for achieving the diversity gain.

2.2. PLC Channel Model

Power line channel is the time-and-frequency variant and exhibits remarkable difference between locations, according to its network topology, the types of wire lines. Moreover, many electrical appliances frequently cause man-made electromagnetic noise on power line channels. Such man-made noise has the impulsive characteristics. Here, we introduce Middleton’s class A noise model [7] into a statistical model of impulsive noise environment. Middleton’s noise model is composed of sum of Gaussian noise and impulsive noise. According to the class A noise model, the PDF (Probability Density Function) of the noise amplitude $\varsigma$ is as follows,

$$p_\varsigma(\varsigma) = \sum_{m=0}^{\infty} \frac{e^{-A A^m}}{m!} \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{\varsigma^2}{2\sigma_m^2}\right),$$

(2)

where $\sigma_m^2 = \sigma^2 (m/\lambda)^{\Gamma} / \lambda^{\Gamma}$, $A$ is the impulsive index, $\Gamma = \sigma_G^2 / \sigma_I^2$ is the GIR (Gaussian-to-impulsive noise ratio) with Gaussian noise power $\sigma_G^2$ and impulsive noise power $\sigma_I^2$, and $\sigma^2 = \sigma_G^2 + \sigma_I^2$ is the total noise power. The noise amplitude $\varsigma$ followed by Eq.(2) always includes the background Gaussian noise with power $\sigma_G^2$. On the other hand, sources of impulsive noise are distributed with Poisson distribution ($e^{-A A^m}/m!$). One impulsive noise source generates noise which is characterized by the Gaussian PDF with variance $\sigma_I^2 / A$. The parameter $A$ is defined as the average number of impulses on the receiver in unit duration times the mean duration of them. As the Gaussian noise power $\sigma_G^2$ is comparatively larger in the total noise power $\sigma^2$, that is, $\Gamma$ is larger, the class A noise will approach to the Gaussian noise. Conversely, the smaller $\Gamma$, the class A noise will be more impulsive.

2.3. A Class of Unitary Space-Time Signal Constellations

Unitary space-time signal is a matrix, whose rows are transmitted from the transmitted elements and mutually orthogonal to each other in wireless communication systems. Let $C \geq 2$ denotes the size of a unitary signal constellation. We define $\theta_C = 2\pi / C$. For any given integers $\eta_1, \eta_2, \eta_3 \in Z$, we define the following unitary signal constellation of size $C$

$$\nu = \nu(\eta_1, \eta_2, \eta_3) = \{\xi(\eta_1, \eta_2, \eta_3) \mid c \in Z_C\},$$

(3)

where $Z_C = \{0, 1, \cdots, C-1\}$, and, $\xi(\eta_1, \eta_2, \eta_3)$ is given by

$$\xi(\eta_1, \eta_2, \eta_3) = \left(\begin{array}{ccc}
e^{j\eta_1 \theta_C} & 0 & 0 \\
0 & e^{j\eta_1 \theta_C} & 0 \\
\cos(\eta_2 \theta_C) & \sin(\eta_2 \theta_C) & 0 \\
-\sin(\eta_2 \theta_C) & \cos(\eta_2 \theta_C) & 0 \\
\eta_3 \theta_C & e^{j\eta_3 \theta_C} & 0 \\
e^{-j\eta_3 \theta_C} & 0 & \eta_3 \theta_C \end{array}\right)^c$$

(4)

For any given constellation size $C \geq 2$, we will find a unitary signal constellation from the following class

$$\Omega_C \equiv \{\nu(\eta_1, \eta_2, \eta_3) \mid \eta_1, \eta_2, \eta_3 \in Z_C\},$$

(5)

such that the unitary signal constellation has the largest diversity product in the constellation class as Eq. (5). The above unitary signal constellation is built from the parametric form of $2 \times 2$ unitary matrices. The signal constellation as Eq. (4) is called parametric code. It is seen that when $\eta_2 = \eta_3 = 0$ is imposed in the constellation class as Eq. (5), the parametric code as Eq. (4) is exactly the diagonal code in the case $M = 2$ where $M$ is the number of diagonal element. There have been several classes of $2 \times 2$ unitary-space-time constellation which were proposed in the previously [6]. A diagonal code cyclic group code for a general $M$ was introduced. A main difference between the diagonal code and the parametric code is that the diagonal code is in general a non-group signal constellation.

3. PLC-VLC/UFM-OFDM

Here, we employ the diagonal code for achieving a diversity gain in the PLC channel. The data stream is divided into bit sequences that consist of $R \cdot M$ bits, where $R$ and $M$ denote information bits per parallel symbol to be transmitted, and the number of diagonal element. Each $R \cdot M$ bit sequence is mapped into the constellation $\nu(c)$ ($c \in Z_C$) selected from $C = 2^RM$. The constellation of unitary matrix can be written as

$$\text{diag}\{\nu(c)\} = \left[\eta e^{j\theta c}, \cdots, \eta e^{j\theta c}\right], \quad (c \in Z_C)$$

(6)

where $\nu(c)$ is $M \times M$ unitary matrix, diag{·} is the diagonal operator, respectively. For example, in the case of $M = 2$ and $R = 1$, which is equal to BPSK modulation. In this case, one of $2 \times 2$ unitary matrix $\nu(c)$ is assigned.

In a PLC channel, we assume that a propagation channel consists of $L$ discrete paths with different time delays due to the power line. The impulse response $h_p(\tau, t)$ is represented as

$$h_p(\tau, t) = \sum_{l=0}^{L-1} h_l(t) \delta(\tau - \tau_l),$$

(7)
where $h_l$ and $\tau_l$ are the PLC complex channel gain and the time delay of the $l$th propagation path, respectively. Through the VLC channel, the channel transfer function $H(f, t)$ is the Fourier transform of $r\zeta h_v(t)h_p(\tau, t)$ and is given by

$$H(f, t) = \int_0^\infty r\zeta h_v(t)h_p(\tau, t) \exp(-j2\pi ft) d\tau$$

$$= r\zeta \sum_{l=1}^{L-1} h_v(t)h_p(\tau, t) \exp(-j2\pi ft). \tag{8}$$

Observing Eq. (8), we split the diagonal components over the coherence bandwidth, the detected signal can achieve the frequency diversity. In this paper, we employ the diagonal code and split diagonal components of the selected code over the coherence bandwidth. Hereafter, we call this processing as an unitary frequency modulation for PLC-VLC/UFM-OFDM (PLC-VLC/UFM-OFDM). In the PLC-VLC/UFM-OFDM systems, the diagonal components of the selected unitary signal constellation are splitting over the coherence bandwidth and are transmitted to the receiver via PLC and VLC channels. In this case, the received signal $Y(k)$ of the $k$-th subcarrier at receiver side is given by

$$Y(k) = H(k)X(k) + N(k) \quad k = 1, \cdots, K, \tag{9}$$

where $X(k)$ is the split diagonal component of unitary signal constellation over the coherence bandwidth, and $N$ is the impulsive noise and additive white Gaussian noise. After de-splitting of the received signals and channel estimation, the frequency domain signals $Y$ are divided into $M$ bits. Here, we consider the same structure of unitary signal constellation for $M \times M$ as Eq. (6). Each $M$ bit of frequency domain signals is demodulated by ML estimator. The ML decision rule of the signal model as Eq. (8) is given by

$$\hat{\nu} = \arg \min_{\nu} \sum_{k=1}^{K} Y(k) - H(k) + \sum_{c=1}^{C} \left| \operatorname{diag}(\nu) \right|^{2},$$

where $\sum_{c=1}^{C} \left| \operatorname{diag}(\nu) \right|$ is the diagonal component of unitary signal constellation. The neighboring signals of $\sum_{c=1}^{C} \left| \operatorname{diag}(\nu) \right|$ without splitting must have correlated channel responses, however, the split signals over the coherence bandwidth achieve totally different channel responses. It means that PLC-VLC/UFM-OFDM systems can achieve a frequency diversity gain.

4. Computer simulated results

In this section, the system performance of the proposed PLC-VLC/UFM-OFDM system is compared with the conventional VLC/OFDM. Fig. 1 shows the simulation model of UFM-PLC/OFDM for $N_c = 128$ subcarriers over the impulsive noise and multipath PLC & VLC channels. In the transmitter, data stream is serial-to-parallel(S/P) transformed, and the diagonal components of unitary signal constellation are split over the coherence bandwidth. The simulation parameters are listed in Table 1.

![Figure 1: Proposed PLC-VLC/UFM-OFDM system.](image)

<table>
<thead>
<tr>
<th>Data Modulation</th>
<th>BPSK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demodulation</td>
<td>ML detection</td>
</tr>
<tr>
<td>Data rate</td>
<td>4 Msymbol/s</td>
</tr>
<tr>
<td>OFDM Symbol duration</td>
<td>65 µsec</td>
</tr>
<tr>
<td>Number of carriers</td>
<td>128</td>
</tr>
<tr>
<td>Channel</td>
<td>2 path quasi-static</td>
</tr>
<tr>
<td>Maximum path delay</td>
<td>0.5 µsec (PLC), 0.3 µsec (VLC)</td>
</tr>
<tr>
<td>Noise model</td>
<td>Middleton’s class A (A=(0.001, 0.01), $\Gamma=(0.001, 0.01)$)</td>
</tr>
</tbody>
</table>
and 64 (A=0.01, Γ=0.01). PLC/UFM-OFDM and PLC/OFDM with weak noise achieve better BER performance than those of with strong noise.

Fig. 4 shows the BER performance of VLC/OFDM, PLC-VLC/OFDM and PLC/UFM-OFDM with heavy (A=0.001, Γ=0.001) and weak (A=0.01, Γ=0.01) noises. From the simulation result, VLC/OFDM and PLC-VLC/OFDM show the approximately same BER performance under the heavy and weak noises. On the other hand, VLC-PLC/UFM-OFDM with splitting size of 64 shows the superior BER performance for heavily and weakly distributed impulse noise channel. This is because PLC/UFM-OFDM can achieve a diversity with splitting the diagonal components over the coherence bandwidth for a frequency diversity in the PLC channel.

5. Conclusion

We have investigated the performance improvement of PLC-VLC/UFM-OFDM over the multipath and impulsive noised PLC and VLC channels, and evaluated the BER performance. The proposed PLC-VLC/UFM-OFDM system can achieve 5 and 4dB gains compared with VLC/OFDM for BER of $10^{-2}$ under the heavy and weak noises, respectively.

References


Novel Channel Estimation and Compensation schemes for Massive MIMO-OFDM
Masaya Okamura†, Chang-Jun Ahn and Ken-ya Hashimoto

†Graduate School of Engineering, Chiba University, Email:aaya2471@chiba-u.jp

Abstract—Since the number of pilot symbols is proportionally depending on the number of transmit antennas, the total transmission rate of massive MIMO would be degraded. To solve this problem, high time resolution carrier interferometry for MIMO/OFDM (HTRC-MIMO/OFDM) has been proposed. HTRC-MIMO systems use a few pilot symbols. Moreover, channel compensation using virtual pilot signal (VPS) has been proposed. However, the number of pilot signals with HTRC-MIMO is not enough for massive MIMO. Since the conventional VPS method iteratively identifies the channel state information (CSI), the complexity is considerable work and the method has low accuracy. From the simulation results, the BER performance of the proposed scheme achieves 2 ~ 2.5dB gains compared with the conventional method based on Hadamard code in 8 × 8 and 16 × 16 MIMO. Moreover, the throughput performance of the proposed scheme achieves 30 ~ 53% improvement compared with the conventional method based on Hadamard code.

1. Introduction

With the spread of mobile network, wireless communications are indispensable in our life. In Japan, 4th Generation mobile communication systems (4G) such as WiMAX2 (Worldwide Interoperability for Microwave Access 2) [1] and LTE-Advanced (Long Term Evolution-Advanced) [2] have been widely used. On the other hand, a variety of novel techniques has been proposed to put it to practical use for 5th Generation mobile communication systems (5G) recently [3]. Since communication frequency band has been tight in recent years, it is necessary to increase band utilization efficiency for 5G. Therefore, high frequency band usage is widely studied for implementation of 5G. Thus, massive MIMO has been widely studied [4]. At present, a wireless LAN has been established a standard up to 8 × 8 MIMO [5]. Furthermore, massive MIMO is considering to use about 100 antennas. Since the number of pilot symbols is proportionally depending on the number of transmit antennas, the total transmission rate of MIMO system would be degraded because the pilot symbol does not carry any information. To solve this problem, high time resolution carrier interferometry for MIMO/OFDM (HTRC-MIMO/OFDM) [6] has been proposed. HTRC-MIMO system has a good BER performance with a few number of pilot symbols. Moreover, channel compensation using virtual pilot signal (VPS) has been proposed [7] [8] [9]. VPS method compensates for data signals with CSI. However, the number of pilot symbols is half as many as Hadamard code in HTRC-MIMO system. It is not enough for massive MIMO. Since the conventional VPS method iteratively identifies the CSI, the complexity and low estimation accuracy are considerable work. To reduce the number of pilot symbols and complexity, we propose a novel HTRC-MIMO with modified orthogonalization method and VPS. This paper is organized as follows. We show the channel model and iterative identification based on VPS in Section 2. In Section 3, we show the proposed schemes. In Section 4, we show the simulation results. Finally, we describe the conclusion in Section 5.

2. System Model

2.1. Channel model

The propagation channel consists of L discrete paths with different time delays, is assumed. The impulse response between the xth transmit and the yth receive antenna \( h_{x,y}(\tau, t) \) is expressed as

\[
h_{x,y}(\tau, t) = \sum_{l=0}^{L-1} h_{x,y,l}(t) \delta(\tau - \tau_{x,y,l}),
\]

where \( h_{x,y,l} \) is the complex channel gain and \( \tau_{x,y,l} \) is the time delay of the \( l \)th path. \( H_{x,y}(f, t) \) is the channel transfer function and the Fourier transform of \( h_{x,y}(\tau, t) \). Then, the channel transfer function \( H_{x,y}(f, t) \) is given by

\[
H_{x,y}(f, t) = \int_{0}^{\infty} h_{x,y}(\tau, t) e^{-j2\pi f \tau} d\tau = \sum_{l=0}^{L-1} h_{x,y,l}(t) e^{-j2\pi f \tau_{x,y,l}}.
\]

2.2. Conventional identification based on VPS

Firstly, the transmitter sends the data signals with forward error correcting code (FEC). In the receiver, the received data signals are demodulated. After decoding, the data signals are obtained. Next, the receiver generates the replica signals \( \hat{d} \) using decoded data signals. The channel variance \( \Delta H_{x,y}^1(k) \) of the xth transmit antenna, the yth receive antenna, the 4th subcarrier and the first iteration using VPS in MIMO system is expressed as

\[
\Delta H_{x,y}^1(k) = \sum_{i=0}^{N-1} \sum_{j=0}^{X-1} \frac{r_x(k, i) - \hat{H}_{x,y}(k)d_y(k, i)}{\hat{H}_{x,y}(k)d_y(k, i)N_S} + 1,
\]
where $\hat{H}_{n}(k)$ is an initially estimated channel response, $d_{r}(k, i)$ is the transmitted signal, $r_{i}(k, i)$ is the received signal, $X$ is the number of transmit antennas and $N_{T}$ is the number of data symbols with the channel re-identification.

Therefore, the adjusted channel response $\hat{H}_{x,y}^{1}(k)$ of the first iteration is given by

$$\hat{H}_{x,y}^{1}(k) = \Delta H_{x,y}^{1}(k)\hat{H}_{x,y}(k).$$ (4)

With the $\xi$ time iterations, the adjusted channel response can be obtained by,

$$\hat{H}_{x,y}^{\xi}(k) = \Delta H_{x,y}^{\xi-1}(k)\hat{H}_{x,y}^{\xi-1}(k).$$ (5)

To prevent the performance deterioration due to the fast channel variance, the following condition is also considered as

$$\hat{H}_{x,y}^{\xi}(k) = \begin{cases} \frac{\hat{H}_{x,y}^{\xi}(k)}{\left|\hat{H}_{x,y}^{\xi}(k)\right|^2} & \text{if } \left|\hat{H}_{x,y}^{\xi}(k)\right|^2 \geq \zeta, \\ \frac{\left|\hat{H}_{x,y}^{\xi}(k)\right|^2}{\hat{H}_{x,y}^{\xi-1}(k)} & \text{otherwise} \end{cases},$$ (6)

where $\zeta (0 \leq \zeta \leq 1)$ is the threshold. By using this threshold, we prevent the deterioration of BER performance. Here, we choose the threshold $\zeta$ using the simulation results [8]. Since we execute iteratively the identification, the channel compensation is accurately operated. As we mentioned above, the total transmission rate of massive MIMO would be degraded with the orthogonal pilot based channel estimation scheme. Moreover, the conventional VPS method is not suitable for massive MIMO since the method identifies the CSI with iterative detection. Therefore, the complexity is also increased.

3. Proposed channel estimation and compensation

In this section, we propose a novel scheme with HTRCI-MIMO with modified orthogonalization methods and VPS method. Since massive MIMO systems use high frequency band, the communications area should be small. In other words, multipath delay spread is short and desired signals are concentrated in a portion. Therefore, when we can easily identify the CSI on the time domain. Moreover, since the proposed scheme uses the initially estimated CSI using VPS, we can reduce the complexity. In the conventional HTRCI-MIMO system, the transmitted pilot signal $d_{ps,i}(k, p)$ of the $x$th transmit antenna, the $k$th subcarrier and $p$th symbol is given by

$$d_{ps,i}(k, p) = \begin{cases} M_{a,b} W_{k} + M_{a-1,b} W_{k} & \text{for } p \leq \left\lfloor \frac{k}{2} \right\rfloor, \\
0 & \text{otherwise} \end{cases}. \quad (7)$$

where $M_{a,b} = \text{mod}(a, b)$ and $W_{k} = \zeta \frac{\sin(\pi k)}{\sin(\pi k)}$. $0 \leq p \leq N_{p} - 1$, $N_{p}$ is the number of pilot symbols and $\left\lfloor x \right\rfloor$ stands for the integer upper and closer to $x$. Since the impulse responses are not overlap to each other in the time domain, we can separate the received pilot signals in each receive antenna. In this paper, we reduce the number of pilot symbols using HTRCI-MIMO with modified orthogonalization method.

In the novel scheme, the transmitted pilot signal is given by

$$d_{ps,n}(k, p) = \begin{cases} \frac{M_{a,b} W_{k+1} + M_{a+1,b} W_{k} + M_{a-1,b} W_{k} + M_{a+1,b} W_{k}'}{4} & \text{for } \left\lfloor \frac{k}{2} \right\rfloor = \text{odd and } \left\lfloor \frac{k}{2} \right\rfloor - 1 \leq p \leq \left\lfloor \frac{k}{2} \right\rfloor \\
(-1)^p \left( M_{a,b} W_{k+1} + M_{a-1,b} W_{k} + M_{a+1,b} W_{k}' \right) & \text{for } \left\lfloor \frac{k}{2} \right\rfloor = \text{even and } \left\lfloor \frac{k}{2} \right\rfloor - 2 \leq p \leq \left\lfloor \frac{k}{2} \right\rfloor - 1 \\
0 & \text{otherwise} \end{cases}. \quad (8)$$

where $\left\lfloor x \right\rfloor$ stands for the integer lower and closer to $x$ and $W_{k}' = \zeta \frac{\sin(\pi k)}{\sin(\pi k)}$. Fig.1 shows pilot signal on the time domain. The figure shows the number of pilot symbols is quarter as many as Hadamard code. The received impulse responses are multiplexed without interference because the transmitted pilot signals are orthogonalized by Hadamard code. Therefore, the received impulse responses are given by

$$h_{r}(p) = \begin{cases} \sum_{n=0}^{\left\lfloor \frac{k}{2} \right\rfloor / 2} \left( -1 \right)^{n} \left\lfloor \frac{k}{2} \right\rfloor / 2 h_{2n+1,y} & \text{for } t = [0, T_{g} - 1] \text{ and } \left\lfloor \frac{k}{2} \right\rfloor - 1 \leq p \leq \left\lfloor \frac{k}{2} \right\rfloor \\
\sum_{n=0}^{\left\lfloor \frac{k}{2} \right\rfloor / 2} \left( -1 \right)^{n} \left\lfloor \frac{k}{2} \right\rfloor / 2 h_{2n+1,y} & \text{for } t = [T_{g}, 2T_{g} - 1] \text{ and } \left\lfloor \frac{k}{2} \right\rfloor - 1 \leq p \leq \left\lfloor \frac{k}{2} \right\rfloor \\
\sum_{n=0}^{\left\lfloor \frac{k}{2} \right\rfloor / 2} \left( -1 \right)^{n} \left\lfloor \frac{k}{2} \right\rfloor / 2 h_{2n+1,y} & \text{for } t = [2T_{g}, 3T_{g} - 1] \text{ and } \left\lfloor \frac{k}{2} \right\rfloor - 1 \leq p \leq \left\lfloor \frac{k}{2} \right\rfloor \\
\sum_{n=0}^{\left\lfloor \frac{k}{2} \right\rfloor / 2} \left( -1 \right)^{n} \left\lfloor \frac{k}{2} \right\rfloor / 2 h_{2n+1,y} & \text{for } t = [3T_{g}, T_{g} - 1] \text{ and } \left\lfloor \frac{k}{2} \right\rfloor - 1 \leq p \leq \left\lfloor \frac{k}{2} \right\rfloor \\
0 & \text{for } \text{otherwise} \end{cases}. \quad (9)$$

where $T_{g}$ is the guard interval (GI) length and $T_{s}$ is the effective symbol length. We assume $T_{s} = \frac{T_{g}}{4}$ in this paper.
Therefore, we can separate these signals into each transmitted signal. Since the power of pilot signals is half of orthogonal pilot signal based on Hadamard code, the channel response is highly affected by the noise. Meanwhile, since we separate four portions of channel responses, we can averaging to reduce the noise effect. After the pilot signal separation, we operate the same procedure as the conventional HTRCI-MIMO. Fig.2 shows the concept of proposed noise whitening. Firstly, we use the initially estimated CSI ($\tilde{h}_{x,y}$) and first estimated CSI using VPS ($\hat{H}_{x,y}$). By applying the IFFT operation, we obtain the time signals. Next, we extract the impulse response of former part ($t = [0, T_s - 1]$). After that, we calculate the Euclidean distance of each impulse response and normalize the distance. Since multipath delay spread is small, the Euclidean distances of the desired impulse responses differ from other responses. In other words, the Euclidean distances of the desired response are small. To eliminate noise component, we insert null signals using threshold level. Moreover, other part ($t = [T_s, T_s - 1]$) is not contained desired impulse response. Therefore, we insert null signals in the part. Finally, we apply the FFT operation to obtain frequency responses. By using these schemes, we can obtain the improved CSI ($h'_{x,y}$).

4. Simulation Results

In this simulation, the modulation is QPSK, the detection method is ZF, FEC is the convolutional code ($R_c = \frac{1}{2}$, $L_c = 7$), the number of antennas, pilot symbols and data are 8 or 16, 2 or 4 and 20, FFT size is 64, the number of subcarriers is 62, GI is 16 sample times, the channel model is 5 path Rayleigh fading and Doppler frequency is 10Hz. Fig.3 shows the simulation result to choose the optimum threshold as 0.15 and 0.10 for 8 × 8 and 16 × 16 MIMO. Fig.4 and Fig.5 show the BER performance for 8 × 8 and 16 × 16 MIMO. As the results, the proposed scheme achieves 2 ∼ 2.5dB gains compared with the orthogonal pilot based channel estimation scheme. Moreover, the BER performance of the proposed scheme is 1.5 ∼ 2dB better than the second iteration. Thus, the proposed scheme achieves good BER performance and low complexity compared with the conventional VPS method. Fig.6 and Fig.7 show the throughput performance for 8 × 8 and 16 × 16 MIMO. From the simulation results, the proposed scheme achieves 30 ∼ 53% throughput compared with the orthogonal pilot based channel estimation scheme.

5. Conclusion

In this paper, we have proposed a novel scheme with HTRCI-MIMO and VPS to improve BER and throughput performance and mitigate the complexity of channel compensation. From the simulation results, it is shown that the proposed scheme achieves 2 ∼ 2.5dB gains compared with the orthogonal pilot based channel estimation scheme for 8 × 8 and 16 × 16 MIMO at BER=10^{-5}. Moreover, the proposed scheme achieves 30 ∼ 53% improvement compared with the orthogonal pilot based channel estimation scheme.
BER vs. Eb/No for 8 × 8 MIMO

BER vs. Eb/No for 16 × 16 MIMO

Throughput vs. Eb/No for 8 × 8 MIMO

Throughput vs. Eb/No for 16 × 16 MIMO

References


Packet Splitting and Adaptive Modulation Based on Time Domain CSI for Cooperative OFDM Systems

Yuta Ida†, Takahiro Matsumoto, and Shinya Matsufuji

Graduate School of Sciences and Technology for Innovation, Yamaguchi University
2–16–1 Tokiwadai, Ube-shi, Yamaguchi, 755-8611 Japan
Email: y.ida@yamaguchi-u.ac.jp†

Abstract—Cooperative communications obtain the transmission and channel diversity gains by using the relay node. However, since the redundancy signal is transmitted to obtain the transmission diversity gain, the transmission rate is degraded. Moreover, since cooperative communications include the interference in the relay node, the gain is degraded. The packet splitting has been proposed based on the channel state information (CSI) of the time domain to obtain the good system performance. In this paper, we propose the combination method with the packet splitting and the adaptive modulation based on the CSI of the time domain to improve the bit error rate (BER) and throughput performances for cooperative orthogonal frequency division multiplexing (OFDM) systems.

1. Introduction

The diversity gain technique is very important to improve the system performance and has been proposed in many fields [1], [2]. In multiple-input multiple-output (MIMO) systems, the transmission and channel diversity gains are obtained by using several transmitted antennas [1]. Similarly, cooperative communications obtain also these diversity gains by using the relay node since there are three links between the source and relay (SR), source and destination (SD), and relay and destination (RD) nodes [2]. On the other hand, cooperative communications degrade the system performance compared with MIMO systems since they include the interference in the relay node. Therefore, in cooperative communications, the relay method is important. In [2], the relay method has been proposed and this paper adapts the decode-and-forward (DF) which is detected the received signal in the relay node and is transmitted to the destination node.

To obtain the transmission diversity gain, the conventional cooperative communications transmit the same signal to the SD and SR links. On the other hand, since this operation degrades the transmission rate, the packet splitting has been proposed [3]-[5]. The packet splitting divides a packet based on the channel state information (CSI) and obtains the good channel diversity gain and prevents a burst error. [4] and [5] have proposed the packet splitting for cooperative communications, but [5] requires many feedback information (FBI) since it needs the CSI for each subcarrier channel. To overcome this problem, [4] have proposed the simple the packet splitting for cooperative communications which divides a packet based on the CSI of the time domain. Since the CSI of the time domain becomes a constant in the case of a small Doppler frequency, the number of FBI becomes only one. Moreover, the adaptive modulation has been proposed to improve the transmission rate and to keep the BER performance [5], [6]. The adaptive modulation changes the modulation level based on the CSI. However, [5] and [6] require also the CSI of each subcarrier channel and many FBI. Therefore, in this paper, we propose the combination method with the packet splitting and the adaptive modulation based on the CSI of the time domain for cooperative orthogonal frequency division multiplexing (OFDM) systems. This paper explains the cooperative OFDM systems in Section II. Then, the proposed packet splitting and adaptive modulation presented, and the computer simulation results are shown in Section III and IV. Finally, the conclusion is given in Section V.

2. Cooperative OFDM Systems

2.1. Channel Model

In the L discrete paths propagation channel with the different delay time, we assume the channel impulse response for the kth node as

\[ h_k(\tau) = \sum_{l=0}^{L-1} h_{k,l} \delta(\tau - \tau_{k,l}). \] (1)

where \( h_{k,l} \) is the complex channel gain which satisfies \( \sum_{l=0}^{L-1} E|\hat{h}_{k,l}|^2 = 1 \), \( E \cdot \cdot \cdot \) is the ensemble average operation, and \( \tau_{k,l} \) is the delay time. After the fast Fourier transform (FFT) operation, the channel response is obtained as

\[ H_k(f) = \int_{0}^{\infty} h_k(\tau) \exp(-j2\pi f \tau) d\tau \]
\[ = \sum_{l=0}^{L-1} h_{k,l} \exp(-j2\pi f \tau_{k,l}). \] (2)

2.2. Cooperative OFDM Systems

Figure. 1(a) shows the structure of the source node. The binary data signal is generated, and the channel coding with
the interleaving is performed. After the serial-to-parallel (S/P) conversion, the coded signal is modulated by the adaptive modulation which will be explain in Section 3. The time domain transmitted signal matrix is given by using the inverse FFT (IFFT) operation as

$$x_k = \sum_{i=0}^{N-1} \sqrt{\frac{S}{N_c}} X_i \cdot w^{-1}, k = \begin{cases} 0 & \text{for } SD, \\ 1,2, \ldots, K & \text{for } SR, \end{cases}$$

where $S$ is the average transmission power, $N_c$ is the number of subcarriers, $X_i$ is the frequency transmitted signal matrix for the $i$th symbol after the adaptive modulation, $w$ is the FFT matrix, $(\cdot)^{-1}$ is the inverse operation, and $K$ is the number of relay nodes. $N_c$ is the number of symbols after the packet splitting which will also be explain in Section 3. After the guard interval (GI) insertion and the parallel-to-serial (P/S) conversion, the time domain signal $x_k$ is transmitted to the relay and destination nodes.

Fig. 1(b) shows the structure of the relay node. The time domain received signal matrix for the $k$th relay node is given by

$$y_k = h_k x_k + z_k,$$

where $h_k$ is the frequency domain received signal matrix for the $k$th symbol after the adaptive modulation, $z_k$ is the additive white Gaussian noise. After the $S/P$ conversion and the GI elimination, the frequency domain received signal matrix $\tilde{X}_k$ is given by

$$\tilde{X}_k = H^{-1}_d Y_k.$$

The equalized signal matrix $\tilde{X}_k$ is converted to the time domain signal matrix $\tilde{x}_k$ as shown in Eq. (3) and is transmitted to the destination node.

Fig. 1(c) shows the structure of the destination node and its time domain received signal matrix is given by

$$y = h_0 x_0 + \sum_{i=1}^{K} h_i \tilde{x}_i + z = \sum_{i=0}^{K} h_i x_i + \tilde{z}_i,$$

where $\tilde{z}$ and $z$ are the AWGN matrices with and without the detection error of the relay node, respectively. The time domain received signal $y$ is processed as shown in Eqs. (5) and (6) as

$$Y_i = \sqrt{\frac{S}{N_c}} y w = \sqrt{\frac{S}{N_c}} \sum_{k=0}^{K} H_{i,k} X_{k,i} + \tilde{Z}_i,$$

and

$$\hat{X}_i = H^{-1}_d Y_i,$$

where $\tilde{Z}_i$ is the AWGN matrix. After the demodulation and the P/S conversion, the detected signal $\hat{X}_i$ is returned to the binary data signal.

3. Proposed Packet Splitting and Adaptive Modulation

3.1. Packet Splitting

For the case of a small Doppler frequency, the power for the CSI of time domain becomes a constant. By using this characteristic, we define its power for $SD$, $SR$, and $RD$ links as $\lambda_{SD}$, $\lambda_{SR}$, and $\lambda_{RD}$. Firstly, the proposed method combines $\lambda_{SR}$ and $\lambda_{RD}$ to consider simply as

$$\lambda_k = (1 - \alpha) \lambda_{SR} + \alpha \lambda_{RD},$$

where $\alpha$ is the weight between $SR$ and $RD$ links. Next, the proposed method decides the number of symbols $N_k$ for Eq. (3). The proposed method gives the number of symbols for $SD$ link as

$$N_0 = \sum_{\epsilon=1}^{K} \frac{N}{\omega} \left[ \frac{\lambda_{SD}}{\lambda_{SD} + \lambda_{\epsilon}} \right],$$

where $N$ is the total number of symbols as $\sum_{\epsilon=0}^{K} N_\epsilon$ and $\omega$ is the weight between $SD$ and $SR$ to $RD$ links, and $[\zeta]$ stands for the integer lower and closer to $\zeta$. Moreover, the number of symbols for $SR$ and $RD$ links is given by

$$N_k = \sum_{\epsilon=0}^{K} \sum_{i=1}^{N_{\epsilon}} \left( N - N_{\epsilon} \right) \left| \frac{\lambda_k}{\lambda_{\epsilon}} \right|,$$

where $K2$ is the number of decided symbols.

3.2. Adaptive Modulation

In this paper, the adaptive modulation for the proposed method consists of a quadrature phase shift keying (QPSK),
a 8 phase shit keying (8PSK), and a 16 quadrature amplitude modulation (16QAM). Moreover, since it is achieved by using the same CSI for the packet splitting, the additional FBI is not required. For the SD link, the proposed method decides the modulation level of the frequency transmitted signal matrix \( \mathbf{X} \) for Eq. (3) as

\[
\mathbf{X}(i) = \begin{cases} 
\mathbf{X}_{4,4th} & \text{for } T_{th1} \geq \lambda_{SD} \\
\mathbf{X}_{3,3rd} & \text{for } T_{th2} \leq \lambda_{SD} < T_{th1} \\
\mathbf{X}_{2,2nd} & \text{otherwise},
\end{cases}
\]

where \( T_{th1} \) and \( T_{th2} \) are the thresholds between a 16QAM and a 8PSK, a 8PSK and a QPSK, respectively, and \( \mathbf{X}_{4,4th}, \mathbf{X}_{3,3rd}, \) and \( \mathbf{X}_{2,2nd} \) are the modulated signal matrices for a 16QAM, a 8PSK, and a QPSK, respectively. Similarly, the adaptive modulated signal matrix for the SR link is also given by

\[
\mathbf{X}(i) = \begin{cases} 
\mathbf{X}_{4,4th} & \text{for } T_{th1} \geq \lambda_k \\
\mathbf{X}_{3,3rd} & \text{for } T_{th2} \leq \lambda_k < T_{th1} \\
\mathbf{X}_{2,2nd} & \text{otherwise}.
\end{cases}
\]

### 4. Computer Simulation Results

In this section, we evaluate the system performance of the proposed method by using the computer simulation. In this simulation, the system model and the parameter are shown in Figure 1 and Table 1. In the source node, the original bit data is generated and is coded by the convolutional code and is interleaved. After the S/P conversion, the parallel signal is modulated. The proposed method adapts the adaptive modulation which consists of a QPSK, a 8PSK, and a 16QAM based on Eqs. (13) and (14). The modulated signal converts the time domain signal by the IFFT operation, and its signal inserts the GI and converts the P/S mode. Here, the proposed method adapts also the packet splitting based on Eqs. (11) and (12). From [4], we define the weights \( \alpha \) and \( \omega \) as \((\alpha, \omega) = (0.3, 1)\). The time domain signal after the packet splitting is transmitted to the two relay and one destination nodes via a propagation channel with a 7 paths Rayleigh fading. In the relay node, the received signal converts the S/P mode and eliminates the GI. The time domain received signal which combines the divided signal for the packet splitting is converted to the frequency domain signal by the FFT operation and is detected by the ZF. The detected signal is demodulated based on the adaptive modulation and converts the P/S mode. After the deinterleaving, its signal is decoded by the Viterbi decoding algorithm and the bit signal is output.

The throughput is defined as

\[
T_p = \frac{N \cdot CR}{T} \left(1 - P_{err}\right),
\]

where \( C \) is the modulation level, \( T \) is the symbol duration, and \( P_{err} \) is the packet error rate (PER). From Eq. (15) and Tab. 1, the maximum throughput for a QPSK, a 8PSK, and a 16QAM is 16, 24, 32 Mbps, respectively. Tab. 2 shows the relation of the throughput between \( T_{th1} \) dB and \( T_{th2} \) dB in \( E_b/N_0 = 25 \) dB. The value of Tab. 2 is derived by using the computer simulation. This means that the value of the throughput is changed by selecting the different threshold. This paper adapts the threshold as \((T_{th1}, T_{th2}) = (0, -5)\) and \((30, 0)\). \((T_{th1}, T_{th2}) = (30, 0)\) means the middle value of the maximum throughput between a QPSK and a 8PSK, and \((T_{th1}, T_{th2}) = (0, -5)\) means also the middle value of the maximum throughput between a 8PSK and a 16QAM.

Fig. 2 shows the throughput versus \( E_b/N_0 \) for the packet splitting and the proposed method. The QPSK, 8PSK, and 16QAM with the packet splitting show the good throughput performance in \( E_b/N_0 < 14 \), \( 14 \leq E_b/N_0 < 17 \), and \( E_b/N_0 \geq 17 \), respectively. The proposed method for \((T_{th1}, T_{th2}) = (30, 0)\) and \((0, -5)\) shows the middle throughput performance between the QPSK and the 8PSK, the 8PSK and the 16QAM, respectively. This is because the proposed method selects the middle value of the throughput from Tab. 2. For the maximum throughput, the proposed method adapts the threshold as \((T_{th1}, T_{th2}) = (30, 0)\) and \((0, -5)\) when the value of \( E_b/N_0 \) is close to the threshold between \( T_{th1} \) dB and \( T_{th2} \) dB.

| Number of subcarriers | 64 |
| Number of relay nodes | 2 |
| FFT size | 64 |
| Guard interval | 16 samples times |
| Path model | 7 paths Rayleigh fading |
| Channel estimation | Perfect |
| FEC | Convolutional code \((R = 1/2, \mathcal{K} = 7)\) |
| Weight | \((\alpha, \omega) = (0.3, 1)\) |
method for \((Th_1, Th_2) = (30, 0)\) and \((0, -5)\) improves about 25\% compared with the QPSK and the 8PSK, respectively.

Fig. 3 shows the BER versus \(E_b/N_0\) for the QPSK (SISO), the packet splitting, and the proposed method. The BER of the QPSK with the packet splitting shows about 2 dB gain compared with the QPSK for SISO since the packet splitting obtains the good channel diversity gain and prevents a burst error. The BER of the proposed method for \((Th_1, Th_2) = (30, 0)\) shows about 1 dB gain compared with the QPSK for SISO and the 8PSK with the packet splitting. This is because the proposed method for \((Th_1, Th_2) = (30, 0)\) has the effect of the QPSK with the packet splitting strongly. Similarly, the BER of the proposed method for \((Th_1, Th_2) = (0, -5)\) shows also about 1 dB gain compared with the 16QAM with the packet splitting since it has the effect of the 8PSK with the packet splitting strongly.

5. Conclusion

In this paper, we have proposed the combination method with the packet splitting and the adaptive modulation based on the CSI of the time domain for cooperative OFDM systems. The proposed method achieves the packet splitting to improve the BER performance and the adaptive modulation to improve the transmission rate based on the same CSI. From the computer simulation results, the proposed method has shown the good BER and throughput performances by adjusting the threshold \(Th_1\) and \(Th_2\).

Acknowledgments

This work was supported by the Mazda Foundation.

References


Analysis of Antenna Pattern Change caused by Architectures close by Radio Transmitting Station

Kazuhsa Haeiwa†, Kazuhiko Kitayama‡, Takeshi Kamio‡, and Hisato Fujisaka‡

†Faculty of Information Science, Hiroshima City University, 731-3194 Japan
‡NHK Integrated Technology Kansai Branch, Osaka 540-0012 Japan
Email: haeiwa-kazu@com.home.ne.jp

Abstract—In medium wave broadcasting stations, architectures close by transmitting station affect antenna pattern and transmitter load impedance. In this paper, the authors show this change of antenna pattern can be analyzed exactly (errors: less than 0.1dB) by considering shape of architectures as cylindrical antennas with same height and area, also introducing radiation efficiency coefficients representing imperfect ground conditions.

1. Introduction

In medium wave broadcasting stations, architectures close by transmitting station are considered as vertical antenna elements having no feeding power. Therefore, mutual coupling among them cannot be neglected, which affects antenna pattern and transmitter load impedance [1]. If influence of those architectures close by transmitting stations could be predicted exactly, it will be very effective to prevent degradation of reception services and shut down of transmitters.

These architectures can be considered to be grounded monopole antennas with same height and area. However, it is impossible to employ conventional antenna theories to this analysis since propagation model for medium waves includes actual ground. Therefore, the authors introduce radiation efficiency representing affection of power losses, which make it possible to apply conventional theories assuming free space for analysis. This proposed method has much more accuracy than conventional ones [2]-[6] (peak searching method) because numerical values are obtained directly.

In this paper, we show that this change of antenna patterns can be analyzed exactly (errors: less than 0.1dB) by considering shape of architectures (26 sites with a height of 20m and more within a distance of 2l) as cylindrical antenna with the same height and area, also introducing radiation efficiency coefficients representing imperfect ground conditions.

2. Formulas Representing Antenna Current

In this paper, the authors employ Hallen’s method [7], [8] for calculating antenna current, which is considered to be most accurate. Figure 1 shows a model of mutual coupling at multiple antenna system composed of monopole antennas. Letting effective feeding voltage (antenna base voltage dominating antenna current: shown in Fig. 1) be $V_T$, antenna current $I(z)$ at $V_T$ is expressed by following formula.

$$I(z) = \frac{jV_T}{60\omega} \sin(\beta L - \vert z \vert) + \frac{b_1}{\omega} \cos(\beta L) + \frac{d_1}{\omega}$$

(1)

Where, $\beta = 2\pi/\lambda$ ($\lambda$: wave length), $L$: antenna height (1/2 of dipole antenna), $z$: antenna vertical position ($0 \leq z \leq L$), and $w$ is expressed by following formula.

$$w = 2 \log \frac{2l}{a}$$

(2)

Where, $a$ is the radius of antennas and $b_1$, $d_1$ are as follows [7].

$$b_1 = F_1(z) \sin(\beta L) - F_1(L) \sin(\beta z) + G_1(L) \cos(\beta z) - G_1(z) \cos(\beta L),$$

(3)

$$d_1 = F_1(L)$$

(4)

Here, letting $F_{oc} = \cos(\beta z) - \cos(\beta L)$, $G_{oc} = \sin(\beta z) - \sin(\beta L)$, formulas shown below are obtained.

$$F_1(z) = -F_{oc} \log(1 - \frac{z^2}{L^2}) - F_{oc} \delta$$

(5)

$$F_1(L) = -\int_{-L}^{l} F_{oc} \exp(-j brA) rA dA$$

(6)

$$G_1(z) = -G_{oc} \log(1 - \frac{z^2}{L^2}) - G_{oc} \delta$$

(7)

$$G_1(L) = -\int_{-L}^{l} G_{oc} \exp(-j brA) rA dA$$

(8)

Meanwhile, $\delta$ is defined by following formula.

$$\delta = \log \left( \frac{1}{4} \left( \sqrt{1 + \left( \frac{a}{L - z} \right)^2} + 1 \right) \left( \sqrt{1 + \left( \frac{a}{L + z} \right)^2} + 1 \right) \right)$$

(9)

Letting antenna current be $I(z)$ with a $V_T$ of 1V, $I(z)$ at $V_T$ is expressed by following formula.

$$I(z) = V_T J(z)$$

(10)

Additionally, antenna impedance (antenna base impedance: $Z_a$) is expressed by following formula.

$$Z_a = \frac{1}{J(0)}$$

(11)
3. Calculation of Wave Level of Each Antenna Using Antenna Pattern Measured Values

Receiving point voltage far from transmitting station $E_j$ is expressed by following formula [7].

$$E_j = \sum_{i=0}^{n} \int_{-L_i}^{L_i} \frac{j60\pi I_i(z)}{\sqrt{\pi}} \exp \left(-\frac{j2\pi r_{ij}}{\lambda} \right) dz$$  \hspace{1cm} (12)

Where, $n+1$: total antenna number (including architectures), $L_i$: each antenna height, $r_{ij}$: distance from the position $(d_i, \phi_i)$ of antennas $A_i$ to a receiving point $(d_j, \phi_j)$. Here, $r_{ij}$ is as follows.

$$r_{ij} = \sqrt{(d_j \sin \phi_j - d_i \sin \phi_i)^2 + (d_j \cos \phi_j - d_i \cos \phi_i)^2}$$  \hspace{1cm} (13)

Letting average antenna current be $I_{ia\text{av}}$.

$$\int_{0}^{L_i} I_i(z)dz = I_{ia\text{av}}L_i = I_{ai}Z_{ai}J_{a\text{av}}L_i$$  \hspace{1cm} (14)

Where, $I_{ai}Z_{ai}$: effective feeding voltage (self feeding voltage-mutual coupling voltage). $J_{a\text{av}}$: average antenna current when effective feeding voltage is IV. Letting $M_i$ be following formula,

$$M_i = 2(I_{ai}Z_{ai}J_{a\text{av}}L_i \sqrt{\eta_i})$$  \hspace{1cm} (15)

Formula [8] is expressed as follows.

$$E_j = \frac{j60\pi}{\lambda} \sum_{i=0}^{n} M_i \exp \left(-\frac{j2\pi r_{ij}}{\lambda} \right) \frac{1}{r_{ij}}$$  \hspace{1cm} (16)

Relation between $E_j$ and $M_i/r_{ij}$ in formula (16) corresponds to a pair of Fourier transform ($\lambda$ is sampling interval). In conventional methods [2]-[6], peak values are searched when $E_j$ is inverse Fourier transformed. In our proposed method, numerical values are obtained directly. Since $E_j$ (measured value) and $r_{ij}$ are known values, $M_i$ ($|M_i|$: wave source level) can be calculated by Kramer’s rule.

4. Calculation of Antenna Pattern

4.1. Formulas representing Antenna Pattern

Mutual coupling voltage $V_{ki}$ at each wave source point ($A_k$ shown in Fig. 1) introducing radiation efficiency is represented as follows [7].

$$V_{ki} = \int_{-L_k}^{L_k} \int_{-L_i}^{L_i} \left(\frac{j60\pi I_i(z)}{\sqrt{\pi}} \right) \frac{M_i}{\sqrt{\eta_i}}$$

\[ \frac{1}{\lambda} \left( 1 + \frac{1}{\beta r_{ik}} \right) \cos \theta_{ik} \]

\[ + \left( \frac{1}{\beta r_{ik}} - \frac{1}{\beta r_{ik}} \right)^2 \sin \theta_{ik} \exp(-j\beta r_{ik}dz)dz \]  \hspace{1cm} (17)

Where, $\beta = 2\pi/\lambda$, $d_{ik}$: distance between $A_k$ and $A_i$, $\cos \theta_{ik} = d_{ik}/r_{ik}$, $r_{ik} = \sqrt{d_{ik}^2 + (z_k - z_i)^2}$. Mutual coupling voltage $V_k$ at point $A_k$ is expressed by following formula.

$$V_k = \sum_{i=0}^{n} V_{ki} + \sum_{i=k+1}^{n} V_{ki}$$  \hspace{1cm} (18)

Letting antenna impedance of wave source $A_k$ be $Z_{ak}$, effective feeding voltage is $I_{ai}Z_{ak}$ (self feeding voltage ($V_{Tk}$) – mutual coupling voltage). Therefore, formula (19) is obtained.

$$I_{ak}Z_{ak} = V_{Tk} - \left( \sum_{i=0}^{k-1} Z_{ai}I_{ai} + \sum_{i=k+1}^{n} Z_{ai}I_{ai} \right)$$  \hspace{1cm} (19)
Formula (20) is obtained using $Z_{kk} = Z_{ak}$.

$$\sum_{i=0}^{n} Z_{ki} I_{ai} = V_{Tk} \tag{20}$$

$I_{ai}$ and $\eta_i$ are obtained by employing Kramer’s rule which means to calculate antenna pattern by using formulas (15) and (16).

### 4.2. Calculation Method

In formula (20), $Z_{ki}$ and $V_{Tk}$ ($V_{Tk} = 0$ except $V_{T0}$) are considered known values, but these are assumed ones actually. However, all parameters including $Z_{ki}$ and $V_{Tk}$ can be obtained by using theories that calculated values and actual ones are equal when calculated antenna pattern is identical with measured one. Therefore, $E_{j(cal)}$ (the results of formula (16) representing antenna pattern) can be obtained. Letting these measured received voltages be $E_{j(meas)}$ and $M$ be total data number, error of measured and calculated ($\Delta$) is expressed by following formula (21).

$$\Delta = \frac{1}{M} \sum_{j=1}^{M-1} \Delta E_j \Delta E_j^* \tag{21}$$

Where, $\Delta E_j = E_{j(cal)} - E_{j(meas)}$, $M$: total data number. $I_{ai}$ and other parameters are obtained by searching points that $\Delta$ becomes minimum (partial differential calculus come to be minimum) by changing all parameters (This one is known as Newton method).

### 5. Comparison of Calculated and Measured Antenna Pattern

In order to confirm usefulness of this method, the authors compared calculated antenna pattern with measured one about a transmitting station having two antennas (main and sub: shown in Fig. 2 with a frequency of 810kHz ($\lambda$: 370.4m). Here, Radiation efficiency of the transmitting station is assumed 1 (100%) because so many radial wires are buried under the ground. Figure. 3 shows a photo of the transmitting station and 26 architectures with a height of 20m and more within a distance of 2\lambda. Figure. 4 shows comparison of calculated and measured antenna pattern. Errors between measured pattern and calculated one are within 0.1dB (0.09dB), which shows this proposed method is very high accurate. Additionally, direction interval of calculated pattern is 1°. Positions, shapes and wave source level of each wave source (transmitting station and architectures) used for simulation are shown in Table 1. Minimum distance among each architectures is 27.9m (0.08\lambda: between No.9 and No.10). No.0 and No.1 are main transmitting and sub antenna respectively.

Distance and position of architectures (wave sources) are known. Meanwhile, their height and radius are calculated values (they are assumed to be cylindrical except main and sub transmission antenna). Wave levels of architectures with a height of under 20m are below -100dB, which produces same results in case those ones are neglected.

### 6. Conclusions

Accuracy of this proposed method is so much higher than conventional ones that errors between calculated and measured values of antenna pattern are less than 0.1dB. This means changes of antenna pattern and load impedance can be predicted exactly at the time when construction of architectures close by transmitting stations are planned, which is very effective to improve broadcasting reception services and prevent shut down of transmitters.
Figure 4: Comparison of calculated and measured antenna pattern.

Acknowledgment

The authors would like to thank Broadcasting Cultural Fund for studying this research.

References


Influence of Coupling Loop Interference on Equivalent C/N Degradation under SFN in a Ship

Takeshi Kamio†, Kazuhisa Haeiwa†, and Hisato Fujisaka†

†Faculty of Information Sciences, Department of Systems Engineering, Hiroshima City University
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
Email: kamio@hiroshima-cu.ac.jp

Abstract– From the viewpoint of frequency utilization efficiency, SFN (single frequency network) is an essential technology for the digital terrestrial TV broadcasting in Japan. We focus on the on-board SFN re-transmission system developed by one of us. Since this system is influenced by the coupling wave, it is important to evaluate the broadcasting signal quality which is often evaluated by the equivalent C/N degradation. However, the existing method cannot be applied to the system, because it does not consider delay wave components over the guard interval (GI) caused by coupling wave. In this paper, we derive the calculation equations to evaluate the equivalent C/N degradation correctly.

1. Introduction

Since nearly 70% of the land in Japan is mountains, there are many transmitter stations (relay stations) for TV broadcasting. Also, since both analog and digital TV had to be broadcasted until 2011, the frequency for digital TV was extremely restricted. Because of such background, ISDB-T (integrated services digital broadcasting - terrestrial) developed in Japan has adopted OFDM (orthogonal frequency division multiplexing) scheme to realize SFN (single frequency network) in which adjacent transmitter stations can use the same frequency.

In the transition period from analog to digital TV, many SFN-gap fillers were placed at poor reception areas (e.g., shade of mountains or buildings) and wave shielded spaces (e.g., underground shopping areas) to receive digital TV waves. Also, Haeiwa, one of the authors of this paper, and others conducted a first-ever research and investigation using a ship connecting Hiroshima and Matsuyama in order to establish methods for stable on-board radio wave reception and on-board SFN re-transmission during sea travel [1]. Although these methods could drastically improve the reception state of digital TV waves, it was also clarified that the coupling loop interference occurred in the on-board SFN re-transmission. However, the influence of coupling wave upon the broadcasting signal quality, which is often evaluated by the equivalent C/N degradation, has not been theoretically analyzed yet.

There already exists a research on the influence of coupling wave of relay SFN upon the equivalent C/N degradation in digital terrestrial TV broadcasting [2]. However, the existing method [2] does not consider delay wave components over the guard interval (GI) caused by coupling wave. Since the on-board SFN re-transmission has an amplifier and the distance between re-transmitting and receiving antennas is short, the delay wave components over GI do not sufficiently decrease. Therefore, the existing method cannot correctly evaluate the equivalent C/N degradation [3] of the on-board SFN re-transmission.

In this paper, we represented the frequency response of re-transmitted wave based on the delay profile observed in the on-board SFN re-transmission [1]. Also, we formulated the equivalent C/N degradation considering not only the frequency response but also the inter-symbol interference (ISI) and inter-carrier interference (ICI) powers. From the results of numerical experiments with transmission parameters of high definition television (HDTV), we have confirmed that the equivalent C/N degradation evaluated by our method is roughly identical with the measured value. Moreover, it has been found that the existing method cannot correctly evaluate the equivalent C/N degradation when the product of the gain of amplifier and the amplitude ratio of coupling wave to re-transmitted wave becomes larger than 0.875.

2. On-board SFN Transmission System

In this paper, we focus on the on-board SFN re-transmission system [1] which was developed to watch the digital terrestrial TV broadcasting in a ship connecting Hiroshima and Matsuyama during sea travel. This system has several receiving antennas placed outside the cabin. After selecting an appropriate receiving antenna in accordance with the electric field strength, the re-transmitted wave is outputted from the SFN re-transmitter in the cabin.

Fig.1 shows the on-board SFN re-transmission system. A selected receiving antenna receives the desired wave \( V_d(t) \) and the SFN re-transmitter outputs the re-transmitted wave \( V_r(t) \). Since this system uses SFN, the receiving antenna receives not only the desired wave but also the coupling wave. When the gain of amplifier and the feedback function are given by \( G \) and \( H \) respectively, this system can be represented as the model of coupling wave shown in Fig.2. Therefore, the relationship between \( V_d(t) \) and \( V_r(t) \) is given by
Also, when the carrier spacing is $f_0$, the carrier index is $k$, the set of carrier indices is $K$, the amplitude ratio of coupling wave to re-transmitted wave is $r$, the delay time between the desired wave and the coupling wave is $\tau_d$, and the initial phase difference is $\alpha$, the feedback function $H$ is given by

$$H = r \sum_{j=k} G \exp\{j(\alpha - 2\pi f_0 \tau_d)\}. \quad (2)$$

Here, paying attention to only the $k$-th carrier, Eq.(1) can be rewritten as follow:

$$V_{o,k}(t) = \frac{GV_{k}(t)}{1 - Gr \exp\{j(\alpha - 2\pi f_0 \tau_d)\}}. \quad (3)$$

3. Derivation of Equivalent C/N Degradation for on-board SFN re-transmission system

3.1. Definition of Equivalent C/N degradation

The equivalent C/N degradation ($C/N_{Ed}$) is often used to evaluate the broadcasting signal quality. $C/N_{Ed}$ is given by the difference between the equivalent C/N ($C/N_{eq}$) and the reference C/N ($C/N_{ref}$) as follows:

$$C/N_{Ed}[dB] = 10\log(C/N_{eq}) - 10\log(C/N_{ref}). \quad (4)$$

The reference $C/N$ ($C/N_{ref}$) corresponds to the receivable limit BER (bit error rate) when receiving only the direct wave. On the other hand, the equivalent $C/N$ ($C/N_{eq}$) is defined as $C/N$ for the estimated channel (transmission line) which is necessary to obtain the BER which is the same as BER corresponding to $C/N_{ref}$.

3.2. Derivation of Equivalent C/N degradation

The existing method [2] assumes that the delay time $\tau_d$ is smaller than GI length ($T_g$) and represents the frequency response $A(k)$ of re-transmitted wave as follows:

$$A(k) = \frac{1}{\sqrt{1 + (Gr)^2 - 2Gr \cos(\alpha - 2\pi f_0 \tau_d)}}. \quad (5)$$

Also, the existing method calculates the equivalent C/N degradation ($C/N_{Ed}$) using only $A(k)$ given by Eq.(5).

Fig.3 shows the delay profile obtained by practical experiments of the on-board SFN re-transmission system [1]. From this figure, we can find the following. The coupling wave continuously generates delay waves at time intervals of $4.1 \mu s$ ($\tau_d$). The level exponentially decreases with the passage of time. Moreover, delay waves exist not only inside GI but also outside GI. In general, since delay waves over GI generate the inter-symbol interference (ISI) and inter-carrier interference (ICI), $C/N_{Ed}$ increases drastically [4]. Therefore, we have to consider the influence of ISI and ICI generated by delay waves over GI to evaluate $C/N_{Ed}$ correctly. To satisfy the above requirement, we have to obtain both the frequency response $A(k)$ considering delay waves over GI and the interference powers (i.e., $P_{ISI}, P_{ICI}$).

First, to obtain $A(k)$, let us consider the Maclaurin’s expansion of Eq.(3) as follows:
Regarding $GV_{d,k}(t)$ as the direct wave, $V_{d,k}(t)$ can be regarded as the receiving wave for multi-path channel where delay waves are continuously received at time intervals of $\tau_i$. Also, when $Gr$ is smaller than one, the level of delay waves exponentially decreases according as the index $i$ increases. Therefore, it is clear that Eq.(6) can represent the delay profile shown in Fig.3.

Fig.4 shows the demodulation behavior based on the delay profile (or Eq.(6)). Noticing that the delay wave over GI in this figure does not include the focused symbol $m$ for $i\tau_d - T_g$, the frequency response $A(k)$ of re-transmitted wave can be represented as follows:

$$A(k) = \left[1 + \sum_{i=1}^{2N} \left(Gr \cos(\alpha - 2\theta_{k,i}) \right) + \sum_{i=1}^{2N} \left(1 - \frac{i\tau_d - T_g}{T_g} \right) (Gr \cos(\alpha - 2\theta_{k,i})) \right]^2,$$

where $T_g$ is the effective symbol length, $I$ is the number of delay waves inside GI, and $N$ is the number of delay waves outside GI.

Moreover, the interference powers ($P_{isi}$, $P_{ici}$) are given by the following equations [5]:

$$P_{isi} = \sum_{i=1}^{I} \left(Gr \frac{i\tau_d - T_g}{T_g} \right)^2,$$

$$P_{ici} = \sum_{i=1}^{I-1} \left(2 - \frac{i\tau_d - T_g}{T_g} \right) \left(1 - \frac{i\tau_d - T_g}{T_g} \right) (Gr)^2.$$

Since there are reports that these interferences are regarded as Gaussian noise from the viewpoint of BER [5], [6], we also obey it. Therefore, $C/N$ of the $k$-th carrier ($C/N_k$) is given by

$$C/N_k = \frac{1}{\sum_{i=0}^{I} \left(\frac{1}{4^k(i/k) - C/N_k} + P_{isi} + P_{ici}\right)}.$$

From the theoretical formula of BER of AWGN channel, the average BER of the on-board SFN re-transmission system is given by substituting $C/N_k$ in Eq.(10) for the following equation:

$$p_r = \frac{a}{N_c} \cdot \sum_{k=1}^{N_c} \text{erfc} \left(\frac{c \cdot (C/N_k)}{b}\right),$$

where, $N_c$ is the number of all carriers, $a$ and $b$ are the coefficients decided by the primary modulation, and $c$ is the coefficient to consider the allocated power of data symbols decreases by pilot symbols. Moreover, the BER corresponding to $C/N_{ed}$ is given by

$$p_{ed} = a \cdot \text{erfc} \left(\frac{c \cdot (C/N_{ed})}{b}\right).$$

From the definition of $C/N_{ed}$, when $pr = p_{ref}$, $C/N_k$ is the equivalent $C/N$. Therefore, substituting $C/N_{ed}$ and $C/N_t$ satisfying $pr = p_{ref}$ for Eq.(4), we can obtain $C/N_{ed}$ of the on-board SFN re-transmission system. On the other hand, the existing method [2] calculates $C/N_{ed}$ in a similar method except that $A(k)$ is given by Eq.(4) and $P_{isi} = P_{ici} = 0$.

### 4. Evaluation of Equivalent C/N Degradation

#### 4.1. Parameters

We carried out the numerical evaluations of equivalent C/N degradation ($C/N_{ed}$) using transmission parameters of HDTV. The transmission parameters and others are summarized in Table 1 [4].

#### 4.2. Comparison of numerically evaluated C/N ED and measured value

Fig.5 shows the number of delay waves vs. $C/N_{ed}$. $Gr$ have been changed from 0.931 to 0.939 at intervals of 0.001. From these results, we have found the following.

If the number of delay waves is smaller than $I = 30$ (i.e., the number of delay waves inside GI), $C/N_{ed}$ is smaller than 5.6dB. Also, if the number of delay waves is larger than $I$, $C/N_{ed}$ increases drastically because the influence of delay waves over GI becomes strong.

Although $C/N_{ed}$ is saturated until $Gr \leq 0.938$, $C/N_{ed}$ is diverged when $Gr = 0.939$. Therefore, it is considered that the on-board SFN re-transmission system is on the edge of oscillation when $Gr = 0.939$. In this case, $C/N_{ed}$ evaluated by our method is 17.1dB.
Fig. 5 The number of delay waves \((i)\) vs. the equivalent C/N degradation \((C/N_{ED})\) when \(C/N_{ref} = 20.1\)dB.

According to the measurements of the actual system [1], the receiving C/N was about 37dB when there was no coupling wave. On the other hand, the receiving C/N corresponding to the edge of oscillation was about 18.5dB under the coupling loop interference. Therefore, the measured \(C/N_{ED}\) was 18.5dB (=37−18.5).

These results mean that \(C/N_{ED}\) evaluated by our method is roughly identical with the measured value.

4.3. Comparison of our method and existing method

Here, we compare our method with the existing method [2] using \(Gr\) vs. \(C/N_{ED}\) as shown in Fig. 6. Both of them have almost the same \(C/N_{ED}\) when \(Gr \leq 0.875\). However, when \(Gr > 0.875\), \(C/N_{ED}\) evaluated by our method increases more rapidly than \(C/N_{ED}\) evaluated by existing method. As mentioned in Sect. 3.2, \(C/N_{ED}\) evaluated by our method is roughly identical with the measured value. Therefore, the existing method cannot correctly evaluate \(C/N_{ED}\) when \(Gr\) becomes large.

5. Conclusions

In this paper, we have formulated the equivalent C/N degradation considering not only the frequency response based on the delay profile but also interference powers to correctly evaluate it in the on-board SFN re-transmission system. As a result, we have confirmed that the equivalent C/N degradation evaluated by our method is roughly identical with the measured value.

References


An improved multi-objective particle swarm optimization using an efficient $G_{\text{best}}$ selection method

Yuki Hasegawa†, Takayuki Kimura‡ and Kenya Jin’no‡

†Graduate School of Electronics, Information and Media Engineering, Nippon Institute of Technology
4-1-1 Gakuendai, Miyasiro, Minami-Saitama, 345–8501, Japan
‡Department of Electrical and Electronic Engineering, Nippon Institute of Technology
4-1-1 Gakuendai, Miyasiro, Minami-Saitama, 345–8501, Japan
Email: hasegawa.2318@gmail.com, tkimura@nit.ac.jp, jinno@nit.ac.jp

Abstract—Accuracy, uniformity and breadth of Pareto-front are important indices for multi-objective optimization problems. However, the conventional multi-objective particle swarm optimization (MOPSO) considers only uniformity to construct the Pareto front. Although the MOPSO shows good performance on the low dimensional multi-objective optimization problems, it shows poor performance for the high dimensional ones. To overcome this problem, we have already improved the MOPSO using determination of initial positions of particles and an effective $G_{\text{best}}$ selection method. In addition, we clarified that the improved method shows better performance than the original MOPSO for the benchmark problems. However, evaluations of our improved method for the real-world problems such as the hybrid renewable energy system is still remaining work. Thus, we evaluate the improved method using the HRES in this study to investigate the applicable possibility of our improved method for the real-world optimization problems. From the numerical experiments, the improved method shows good performance for rated value decision problems in the HRES.

2. Multi-Objective Particle Swarm Optimization (MOPSO)

In the MOPSO[1], $x_i(t)$ is the position of the $i$ th particle at the $k$ th iteration, and $v_i(t)$ is the velocity of the $i$ th particle at the $k$ th iteration. Then, to find the optimum Pareto-front, the MOPSO algorithm works as follows:

step 1 : Initialization of the particles
The initial position of the $i$ th particle, $x_0(i)$, is randomly determined. In addition, the initial velocity of the $i$ th particle, $v_0(i)$, is set to 0.

step 2 : Updating the $P_{\text{best}}(i)$ and $G_{\text{best}}$
$P_{\text{best}}(i)$ is the position of the $i$ th particle where the best evaluation value was obtained in the past searching history. $G_{\text{best}}$ is the position of the particle when the best evaluation value were obtained among all the particles. In the searching process of the MOPSO, if a non-dominated solution is obtained, this solution is stored in the archiving space. $P_{\text{best}}(i)$ is updated if the $i$ th particle finds the best evaluation value. In addition, $G_{\text{best}}$ is also updated if the particle finds the best value during the searching history.

step 3 : Formation of the hypercube
The solution space is divided to form the hypercubes. The hypercube is used to measure the density of the solutions in the solution space.

step 4 : Selection of the global best
To select the global best used in MOPSO, the evaluation value for the hypercube that expresses the density of the solutions in the solution space is defined as follows:

$$D_i = \begin{cases} \frac{\text{rand}(0,1)}{H_i} & \text{(if a solution exists in the } i \text{ th hypercube)}, \\ 0 & \text{(otherwise)}. \end{cases}$$

(2)
In Eq. (2), \( D_i \) is the density of the solutions in the \( i \) th hypercube and \( H_i \) is the number of solutions in the \( i \) th hypercube. If Eq. (2) takes large value, or the small number of solutions exist in a hypercube, one of the solutions in the hypercube is selected as the next global best.

**step 5**: Updating the velocity and position of the particles

\[
v_{k+1}(i) = \omega v_{k}(i) + rand_1 C_1 (P_{best}(i) - x_k(i)) + rand_2 C_2 (G_{best} - x_k(i)),
\]

\[
x_{k+1}(i) = x_k(i) + v_{k+1}(i).
\]

In Eq. (3), \( \omega \) is the inertia weight, \( rand_1 \) and \( rand_2 \) are the uniformly distributed random numbers within the range of \([0, 1]\), and \( C_1 \) and \( C_2 \) are the acceleration weights.

**step 6**: Evaluation of the particles

Calculation of the objective function is finally conducted using the current position of the particles.

Steps 2 to 6 are repeated until the termination condition is satisfied.

3. Our improved method using determination of initial positions of particles and an effective \( G_{best} \) selection method[2]

In this section, we discribed an improved MOPSO method[2].

3.1. Determination of initial position of particle

To eliminate the possibility for getting into the local minima by the MOPSO, we add the determination method for the initial positions of particles[4] into the conventional MOPSO (DP-MOPSO). In the DP-MOPSO, the searching process of the optimum Pareto front is divided into two stages. In the first stage, the optimum points in the solution space are obtained using the single PSO (Fig. 1(a)). Obtained positions of the particles are stored in the archiving space. In addition, the searching process of the optimum point for each objective function is performed in the second stage, and the obtained positions are also stored in the archiving space. The initial positions of the particles are randomly selected from the ones stored in the archiving space. Schematic diagram for the determination method of initial positions of particles in the DP-MOPSO is described in Fig.1.

3.2. An effective \( G_{best} \) selection method

In the conventional MOPSO, the ununiformal Pareto front is obtained because the global best \( G_{best} \) that determines the directions of all the particles using Eq. (2). We then change this global best selection method using improved crowding-distance-assignment. By using the improved crowding-distance-assignment, the solution that is the closest to the optimum solution is easily selected as the \( G_{best} \) at each iteration. The improved crowding-distance-assignment of the \( i \) th non-dominated solution is defined as follows:

\[
CD_i = \frac{1}{EUC(ND_i, Opt)} \frac{ND_{i+1} - ND_{i-1}}{f_{max} - f_{min}}
\]

In Eq. (5), \( EUC(ND_i, Opt) \) is an euclidean distance between the optimum point \( Opt \) and the \( i \) th non-dominated solution \( ND_i \). \( f_{max} \) and \( f_{min} \) is the maximum and minimum values of the objective functions. Schematic diagram of the effective \( G_{best} \) selection method is described in Figs.2.

4. Hybrid Renewable Energy System (HRES)

A HRES[3] is a power supply system integrating the renewable energy systems to reduce the unpredictability of the power by the renewable energies (Fig. 3). In the HRES, the output power to meet the demand of the consumer is mainly produced by the photovoltaic power generation system and the wind turbine. If the output power is greater than the power demand, surplus electric power will be stored in the battery. In addition, if the charging electric power for the battery meets the maximum capacity, remaining power will be used to extract the hydrogen from \( \text{H}_2\text{O} \), then, the hydrogen is charged into the hydrogen tank. If the output power from the photovoltaic system and the wind turbine cannot meet the energy demand, the saving power by the battery and the hydrogen tank will be used to supply the shortage power. Further, if the power from the
Figure 3: Configuration of the hybrid renewable energy systems.

photovoltaic system, the wind turbine, the battery and the hydrogen tank can not meet the demand, the diesel generator starts to supply the power demand as the backup power, then, the CO2 emission from the HRES increases.

The rated power of the system $P$ is defined as follow:

$$P = [P_{PV}, P_{WG}, P_{bat}, P_{EL}, P_{tank}, P_{FC}, P_{DG}],$$

where $P_{PV}$[kW] is the rated power of the photovoltaic power generation system, $P_{WG}$[kW] is the rated power of the wind turbine power generation system, $P_{bat}$[kW] is the rated capacity of the battery, $P_{EL}$[kW] is the rated power of the electrolytic cell, $P_{tank}$[kg] is the rated capacity of the hydrogen tank, $P_{FC}$[kW] is the rated power of the fuel cell, and $P_{DG}$[kW] is the rated power of the diesel generator.

Mathematical model of each system is described as follows:

- **Wind turbine power generation system:**
  
  Output power from the wind turbine power generation system, $E_{WG}$, is defined as follows:

$$E_{WG} = \begin{cases} 0 & v \leq V_c, \\ \frac{3}{8} P_{WG} \left( \frac{V_c}{V_f} \right)^3 & V_c \leq v < V_f, \\ P_{WG} & V_f \leq v < V_f, \\ 0 & v \geq V_f, \end{cases}$$

where $V_c$ is the cut-in wind speed, $V_f$ is the rated wind speed, and $V_f$ is the cut-out wind speed, and $P_{WG}$ is the rated power of the wind turbine.

- **Photovoltaic power generation system:**
  
  Output power from the photovoltaic power generation system, $E_{PV}$, is defined as follows:

$$E_{PV} = \eta_{PV} P_{PV},$$

where $P_{PV}$ is the rated power of the PV panel, and $\eta_{PV}$ is the conversion efficiency.

- **Electrolyzer:**
  The amount of hydrogen $Q_{H_2}$ is defined as follows:

$$Q_{H_2} = \frac{E_{EL}}{\alpha_E \eta_{EL}},$$

where $E_{EL}$ is the power consumption of the electrolyzer, $\alpha_E$ is a coefficient of power consumption curve, and $\eta_{EL}$ is the electrolysis efficiency.

- **Fuel cell:**
  The fuel cell is used as a backup generator by converting electrical energy into chemical energy of hydrogen and an oxidant. Consumption of hydrogen in the fuel cell $H_{2\text{cons FC}}$ is defined by,

$$H_{2\text{cons FC}} = \alpha_{FC} P_{a FC} \eta_{FC},$$

where $P_{a FC}$ is the output power of fuel cell, $\alpha_{FC}$ is a coefficient of hydrogen consumption curve, and $\eta_{FC}$ is a power generation efficiency.

- **Diesel generator:**
  The diesel generator is used as an emergency power source. The fuel consumption of the diesel generator, $fuel_{cons}$, is defined by,

$$fuel_{cons} = \alpha_{DG} P_{a DG} \eta_{DG},$$

where $P_{a DG}$ is the output power of the diesel generator, $\alpha_{DG}$ is a coefficient of fuel consumption curve, and $\eta_{DG}$ is a power generation efficiency.

- **Battery:**
  The state of charge at the $t$th time of the battery, $SOC(t)$, is defined by,

$$SOC(t) = \begin{cases} SOC(t-1) + \frac{S_{SOC(t)} \eta_{charge}}{P_{bat}} \times 100 & \text{(charging)}, \\ SOC(t-1) - \frac{S_{SOC(t)} \eta_{discharge}}{P_{bat}} \times 100 & \text{(discharging)}, \end{cases}$$

$$SOC_{\text{min}} \leq SOC(t) \leq SOC_{\text{max}},$$

where $E_{bat}$ is the amount of power charging or discharging the battery, $P_{bat}$ is the rated capacity of the battery, $\eta_{charge}$ is a charging efficiency, $\eta_{discharge}$ is a discharging efficiency. The battery can discharge the energy until the power reaches $SOC_{\text{min}}$, and charges until it reaches $SOC_{\text{max}}$.

- **Hydrogen tank:**
  The amount of the hydrogen at the $t$th time, $H_{2\text{level}}(t)$, is defined by,

$$H_{2\text{level}}(t) = H_{2\text{(t-1)}} + Q_{H_2}(t) - \frac{H_{2\text{cons FC}}(t)}{\eta_{H_2 - \text{tank}}},$$

$$H_{2\text{level min}} \leq H_{2\text{level}}(t) \leq H_{2\text{level max}},$$
where \( Q_{H2} \) is the mass flow rate of hydrogen electrolyzer, \( H_{2,cons,FC}(t) \) is the hydrogen consumption of the fuel cell, and \( \eta_{H2,tank} \) is the storing efficiency. 

Also, the hydrogen tank has an upper limit and a lower limit to store the hydrogen. The upper limit of the hydrogen tank, \( H_{2,level,max} \), is defined by the rated capacity of hydrogen, and the lower limit, \( H_{2,level,min} \), is defined by 5% of the rated capacity of the hydrogen tank.

To determine the optimum rated value of the HRES, we consider the following objectives, 1) the amount of the \( CO_2 \) emission, and 2) the total cost of the system. The total cost includes the initial cost, the operation, maintenance costs, the replacement cost, and fuel costs. Each objective for the HRES is defined as follows:

- **The amount of the \( CO_2 \) emission**

  The amount of the \( CO_2 \) emission is defined by,

  \[
  CO_{2,\text{emission}} = \sum_{t=1}^{8760} fuel_{cons}(t) \times EF, \quad (16)
  \]

  where \( EF \) is the emission factor of diesel generator. We set the cost of the EF €1.2 for 2.5[kg/Lt].

- **Total cost of the system**

  The total cost of the system is defined as follows:

  \[
  \text{cost} = \sum_j [C_{I,j} + C_{OM,j} \times \frac{1}{CRF(i,T)} + C_{rep,j} K_j P_j] + C_{IS,j} + C_{OMS,j} + C_{fuel} \times fuel_{cons,yr} \times \frac{1}{CRF(i,T)}, \quad (17)
  \]

  \[
  K_j = \sum_{i=1}^{n} \frac{1}{1 + (1 + ir)^{L_{j,i}}}, \quad (18)
  \]

  \[
  Y = \begin{cases} 
  \frac{T}{T} - 1 & T \% L = 0, \\
  \frac{T}{T} & T \% L \neq 0,
  \end{cases} \quad (19)
  \]

  \[
  CRF = \frac{ir(1 + ir)^T}{(1 + ir)^T - 1}, \quad (20)
  \]

  where \( C_{I,j} \) is the investment cost of the component \( j \), \( C_{OM,j} \) is the operation and the maintenance cost of the component \( j \), \( C_{rep,j} \) is the replacement cost of the component \( j \), \( C_{fuel} \) is fuel costs, \( fuel_{cons,yr} \) is the fuel consumption by Lt/yr annual, \( C_{IS,j} \) is the intercept of the investment cost of the component \( j \), \( C_{OMS,j} \) is the intercept of the operation and maintenance cost of the component \( j \), \( K_j \) is the worth of the single payment, \( CRF \) is the capital recovery factor, represented in Eqs. (18), (19) and (20), respectively. In Eqs. (18), (19) and (20), \( ir \) is interest rates, \( L \) and \( Y \) are life time and the number of replacement of the component \( j \), \( T \) is 25-year project life.

5. **Numerical experiments**

In these numerical experiments, we compared the performance of the IC-DPMOPSO with that of the conventional MOPSO for the rated value decision problems. We set the number of iterations to 1,000, the number of particles to 200. Results of the Pareto solution set obtained by the conventional MOPSO and IC-DCMOPSO for the rated value decision problem of HRES are shown in Fig. 4. In Fig. 4, pareto solution set by the IC-DPMOPSO shows better uniformity than that by the conventional MOPSO.

![Figure 4: Pareto solution set obtained by the conventional MOPSO and IC-DCMOPSO](image)

6. **Conclusion**

In this study, we evaluated the improved MOPSO[2] for the rated value decision problem of HRES to investigate the applicable possibility of our improved method for the real-world optimization problems. From the results of the numerical experiments, we confirmed the improvement of uniformity and breadth of the pareto front can be obtained by the improved method. In the future work, we propose a better solution method as compared to improved method[2] for the other real multi objective optimization problems.

The research of T.K. was partially supported by a Grant-in-Aid for Young Scientists (B) from JSPS (No.16K21327).

References


Principal Component Stock Portfolio Based on Nonlinear Prediction

Kazuki Yanagisawa and Tomoya Suzuki

Graduate School of Science and Engineering, Ibaraki University
4-12-1 Nakanarusawa-cho, Hitachi-shi, Ibaraki 316-8511, Japan
Phone/FAX: +81-294-38-5195, E-mail: [15nm926x, tomoya.suzuki.lab]@vc.ibaraki.ac.jp

Abstract—We have verified that the principal-component analysis (PCA) can enhance the predictive power of stock price movements by the statistical significance test (the paired t-test), which might be thanks to the noise reduction effect by the PCA. As its application, we used this concept for our nonlinear portfolio model and confirmed its validity by using real stock data.

1. Introduction

In our previous study, we proposed the nonlinear principal-component portfolio (NCP) model[1] based on a nonlinear prediction method with the principal component analysis (PCA) for stock price movements, and confirmed its validity by using real stock data. That is, the eigenvalue corresponding to the l-th eigenvector \( \lambda_l \) of all stocks. This hidden in them can be extracted. The more detail will be confirmed independently to each principal component \( r_i(t) \), \( i = 1, 2, \ldots, N \) at the time \( t \), and let \( \Sigma(t) \) denote the covariance matrix of historical return rates of all stocks. By diagonalizing \( \Sigma(t) \),

\[
E^{-1} \Sigma(t) E = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)
\]

where \( E = [e_1, e_2, \ldots, e_N] \) is the eigenvector matrix, and \( \lambda_i \) is the eigenvalue corresponding to \( e_i \). The l-th principal component score is calculated by the l-th eigenvector \( e_i \) as follows:

\[
r_i(t) = r(t)e_i,
\]

where \( r(t) = [r_1(t), r_2(t), \ldots, r_N(t)] \). (2)

After that, we apply the nonlinear prediction method used in the previous study[3] independently to each principal component \( r_i(t) \), \( i = 1, 2, \ldots, N \), and obtain each predicted values \( \tilde{r}_i(t+1) = [\tilde{r}_1(t+1), \tilde{r}_2(t+1), \ldots, \tilde{r}_N(t+1)] \). Then, it is inversely transformed by

\[
\hat{r}(t+1) = \tilde{r}(t+1)E^T
\]

2. Principal Component Based Nonlinear Prediction

Let \( r_i(t) \) denote the return rate of i-th stock \( i = 1, 2, \ldots, N \) at the time \( t \), and let \( \Sigma(t) \) denote the covariance matrix of historical return rates of all stocks. By diagonalizing \( \Sigma(t) \),

\[
E^{-1} \Sigma(t) E = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)
\]

where \( E = [e_1, e_2, \ldots, e_N] \) is the eigenvector matrix, and \( \lambda_i \) is the eigenvalue corresponding to \( e_i \). The l-th principal component score is calculated by the l-th eigenvector \( e_i \) as follows:

\[
r_i(t) = r(t)e_i,
\]

where \( r(t) = [r_1(t), r_2(t), \ldots, r_N(t)] \). (2)

After that, we apply the nonlinear prediction method used in the previous study[3] independently to each principal component \( r_i(t) \), \( i = 1, 2, \ldots, N \), and obtain each predicted values \( \tilde{r}_i(t+1) = [\tilde{r}_1(t+1), \tilde{r}_2(t+1), \ldots, \tilde{r}_N(t+1)] \). Then, it is inversely transformed by

\[
\hat{r}(t+1) = \tilde{r}(t+1)E^T
\]

3. Statistical Significance Test

To compare the predictive powers given by Method A and Method B, we perform the paired t-test. Here, we denote the prediction accuracy given by Method A as \( C^A = [C_{11}^A, C_{12}^A, \ldots, C_{NN}^A] \) and that by Method B as \( C^B = [C_{11}^B, C_{12}^B, \ldots, C_{NN}^B] \). If the prediction accuracy of Method A is better than that of Method B, then \( C^A > C^B \). Then, we define \( d_i = C_{ii}^A - C_{ii}^B \), and calculate the test statistic \( T \):

\[
T = \frac{\text{E}[d_i]}{\text{std}(d_i) / \sqrt{N}}
\]

where \( \text{E}[d_i] \) and \( \text{std}(d_i) \) are the mean value and standard deviation of \( d_i \). If \( T \) follows the standard normal distribution \( N(0, 1) \), \( T > 1.676 \) means a significant difference between \( C^A \) and \( C^B \) in the one-sided test at the significance level of 5%. If so, this can be an evidence that the principal-component preprocessing enhanced the predictive power of stock price movements.

For the test, we randomly selected 50 stocks from 500 stocks listed on the Tokyo Stock Exchange. According to Table 1, \( C^A \) is statistically larger than \( C^B \) in all periods except 2008–2010. It can conclude that the predictive power has been increased by the PCA.

4. Conclusions

By the paired t-test, we confirmed that the PCA can improve the prediction accuracy of nonlinear prediction. It might be because the PCA can cancel out random behaviors of all stocks, and therefore the principal nonlinear structure hidden in them can be extracted. The more detail will be discussed in the presentation, especially about the risk reduction power by composing stock portfolios.

Table 1: The test statistic \( T \) in each period

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( T )</td>
<td>8.500</td>
<td>6.592</td>
<td>0.160</td>
<td>2.827</td>
</tr>
</tbody>
</table>

References

Biased Reactions to Abnormal Stock Prices Detected by Autoencoder during Daytime and Nighttime

Hiroyuki Gotou and Tomoya Suzuki

Graduate School of Science and Engineering, Ibaraki University
4–12–1, Nakanarusawa-cho, Hitachi-shi, Ibaraki, Japan
E-mail: [16nm921n, tomoya.suzuki.lab@vc.ibaraki.ac.jp]

Abstract—We applied an autoencoder to learn the relationships among all stocks, and considered the stock prices that the autoencoder cannot restore as abnormal prices. Then, we identified that the price movements following the abnormal prices caused during the nighttime are clearly biased. As its application, we proposed contrarian trading strategies and confirmed each validity by some statistical significance tests.

1. Introduction

The market price of an individual stock sometimes moves differently from that of the other stocks if good or bad news related to only the individual stock is reported. However, such an irregular price movement sometimes leads to the overreaction of traders to the news, and therefore its stock price exceeds the reasonable price. If so, this stock price might be immediately modified into the reasonable price owing to the efficiency of the market.

From this viewpoint, our previous study [1] used an indicator to detect abnormal stock prices and investigated the reaction of financial markets to them. However, abnormal stock prices were detected by comparison with recent historical prices of the same stock, and the relationship with other stocks was not considered. For this reason, our next study [2] focused on the relationships among all stocks and applied an autoencoder [3] to detect the abnormal behavior of individual stocks. However, this study focused only on daily data. For more advanced research, the present study separates daily price movements into shorter timescales: daytime movements and nighttime movements. Because the market is closed during the nighttime, the mechanism of creating abnormal prices and their modifications might be different in each timescale.

The autoencoder is well known as a pretraining tool of deep neural networks and can extract the characteristics of all stocks as nonlinear principal components so as to restore the original stock prices from their principal components. If an individual stock price is different from the output from the learned autoencoder, the price can be considered unusual. If the unusual price was created by the overreaction of traders, the following price movement is more predictable than usual because the reverse movement is expected. To confirm the validity of this concept, we analyze two major stock markets in Japan and USA.

2. Autoencoder

The autoencoder [3] is a kind of neural network with the output layer having the same number of neurons as the input layer. The middle layer has fewer neurons because it is used for dimension reduction to extract the principal characteristics of the input layer. Then, the output layer restores the input information compressed by the middle layer. Namely, the autoencoder learns only the essential structure of the inputted data. Therefore, if the autoencoder after learning cannot restore new input data, then this new data can be considered to be generated by an unlearned structure.

In the present study, we input return rates of many stocks to an autoencoder. If we denote the dealing price of the ith stock at time t as \( x_i(t) \), its return rate \( r_i(t) \) is defined by

\[
r_i(t) = \frac{x_i(t) - x_i(t-1)}{x_i(t-1)}. \tag{1}
\]

The input layer of an autoencoder obtains a set of return rates \( [r_i(t)]_i = 1, 2, \cdots, N \), and then the jth neuron of the middle layer outputs \( y_j(t) \):

\[
y_j(t) = f \left( \sum_{i=1}^{N} w_{ij} r_i(t) + b_j \right), \quad f(u) = \frac{1}{1 + \exp(-u)} \tag{2}
\]

where \( w_{ij} \) is the connection strength, \( b_j \) is a bias term, and \( j = 1, 2, \cdots, J \). In this study, the number of neurons in the middle layer is set as \( J = 0.9N \). Next, the output layer restores the set of \( \{r_i(t)\} \) by

\[
\hat{r}_i(t) = g \left( \sum_{j=1}^{J} w_{ji} y_j(t) + b_i \right), \quad g(u) = u \tag{3}
\]

where \( w_{ji} \neq w_{ij} \).

This autoencoder learns the principal structure among all N stocks using the back-propagation algorithm so as to minimize the total error between the inputs \( \{r_i(t)\} \) and their outputs \( \{\hat{r}_i(t)\} \). Here, if we denote the present time as \( T \), the learning term is \( T - \tau \leq t \leq T - 1 \). In this study, we set \( \tau = 100 \) [days].
The back-propagation algorithm is as follows:

\[ w_{ji} \leftarrow w_{ji} + \eta \delta_j y_j(t), \]  
\[ w_{ij} \leftarrow w_{ij} + \eta \delta_i r_i(t), \]  
\[ b_j \leftarrow b_j + \eta \delta_j, \]  
\[ b_i \leftarrow b_i + \eta \delta_i, \]

where

\[ \delta_i = (r_i - \hat{r}_i)^T f'(u_i), \quad u_i = \sum_{j=1}^{N} w_{ji} y_j(t) + b_i, \]  
\[ \delta_j = \left( \sum_{i=1}^{N} w_{ij} \delta_i \right) g'(u_j), \quad u_j = \sum_{i=1}^{N} w_{ij} r_i(t) + b_j, \]

and \( \eta \) is a learning coefficient. We repeat the above modification as long as the difference between \( |r_i(t)| \) and \( |\hat{r}_i(t)| \) can be reduced. However, to avoid overfitting to the learning data, the newest 10% of the learning data is preserved to evaluate this difference and the other 90% is used for the back-propagation algorithm to modify the model parameters.

Namely, the first 90 days are used for the back-propagation algorithm to train the autoencoder and the next 10 days are used as the test data to evaluate its generalization ability. When the present time \( T \) is updated, the autoencoder is trained again.

3. Detection of Abnormal Stock Prices

3.1. Indicator based on autoencoder

After the autoencoder is trained, we input the newest data set of all stocks \( \{r_i(T)\}_{i=1}^{N} \), and calculate the following indicator:

\[ I_i(T) = \frac{r_i(T) - \hat{r}_i(T)}{\sigma_i} \]  

where \( \{\hat{r}_i(T)\}_{i=1}^{N} \) is the output from the trained autoencoder. Then, \( \sigma_i \) is the standard deviation of the fitting errors of the test data, that is, the standard deviation of \( |r_i(t) - \hat{r}_i(t)|T - 0.1 \leq t \leq T - 1 \).

Next, if \( I_i(T) \) is over 3 or under \(-3\), we consider that \( r_i(T) \) is an unusual return rate. This is because \( 2\sigma \) corresponds to about 5% significance level in the normal distribution. However, in real financial markets, return rates have fatter tails than the normal distribution, and therefore we set \( 3\sigma \).

3.2. Reaction to abnormal stock prices

We applied the indicator \( I_i(T) \) to 617 Japanese stocks listed on Tokyo Stock Exchange in 2000–2006 and to 700 American stocks listed on New York Stock Exchange in 2001–2007. We classified abnormal prices into two types: daytime movements and nighttime movements. The daytime movement is the price difference from the opening price to the closing price in the same day, and its reaction is observed at the next opening price. On the other hand, the nighttime movement is that from the closing price to the opening price in the next business day, and its reaction is observed until the next closing price. Moreover, the abnormal price movement has two directions; \( I(T) \geq 3 \) is a rising movement and \( I(T) \leq -3 \) is a falling movement. If the abnormal price movement detected by the indicator \( I(T) \) does not have a fair reason, the price might be modified soon into the proper price by the efficiency of the market. To confirm this assumption, we calculated some conditional probabilities of the next return rate \( r(T + 1) \). The results are shown in Tables 1 and 2.

As we expected, the reversal of the price tends to occur if \( I(T) \geq 3 \) or \( I(T) \leq -3 \). The conditional probability \( P(r(T + 1)|I(T)) \) is clearly higher than \( P(r(T + 1)|\hat{r}(T)) \) and \( P(r(T + 1)) \). It might be caused by the overreaction of traders and its modification into the appropriate price, which can be considered as a new anomaly in stock markets contradicting the efficient-market hypothesis. In particular,
this biased reaction becomes clearer right after the falling price movement (i.e., \( I(T) \leq -3 \)) than the rising one (i.e., \( I(T) \geq 3 \)). It might be because of the asymmetric human mind; according to the prospect theory, people think lightly of profit but think heavily of getting loss. For this reason, falling price movements easily lead to the overreaction of traders, and then the stock price becomes less than the reasonable level. In addition, it is also important to focus on the difference between daytime and nighttime. The biased reaction is clearer right after the nighttime’s price movement than the daytime’s one. As a reason, because stock markets are closed during the nighttime hours, any traders cannot make a deal even if important news happens. In this case, new orders are reserved through the night, and all of them are concentratedly dealt with at the next opening time. This mechanism causes excessive price movements. However, due to the efficiency of the market, this overreaction is immediately modified into the proper level. These reasons might be why the biased reaction is clearly confirmed right after the nighttime’s falling price movement.

4. Trading Strategy Foreseeing the Biased Reaction

If there is the above anomaly in stock markets, it increases the predictability of future price movements. From this viewpoint, we can consider contrarian trading strategies as follows:

(SaD) We short-sell the \( i \)th stock at time \( T \) right after the daytime’s rising price movement of \( I_i(T) \geq 3 \).

(BaD) We buy the \( i \)th stock at time \( T \) right after the daytime’s falling price movement of \( I_i(T) \leq -3 \).

(SaN) We short-sell the \( i \)th stock at time \( T \) right after the nighttime’s rising price movement of \( I_i(T) \geq 3 \).

(BaN) We buy the \( i \)th stock at time \( T \) right after the nighttime’s falling price movement of \( I_i(T) \leq -3 \).

Then, we immediately close the position at the next time \( T + 1 \). If some stocks satisfy the above condition at the same time \( T \), we allocate all investment money into these stocks at the same rate.

To evaluate the investment performance of each strategy, we performed investment simulations using the new real data of 617 Japanese stocks in 2007–2013 and 700 American stocks in 2008–2014. Bold figure means the best score in each category.


<table>
<thead>
<tr>
<th>Country</th>
<th>Japan</th>
<th>America</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strategy</td>
<td>SaD</td>
<td>BaD</td>
</tr>
<tr>
<td>( F )</td>
<td>1454</td>
<td>1262</td>
</tr>
<tr>
<td>( R[%] )</td>
<td>25.9</td>
<td>186</td>
</tr>
<tr>
<td>( \bar{R}[%] )</td>
<td>0.02</td>
<td>0.15</td>
</tr>
<tr>
<td>( W[%] )</td>
<td>49.2</td>
<td>58.6</td>
</tr>
<tr>
<td>( \phi )</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>( \psi )</td>
<td>1.03</td>
<td>1.36</td>
</tr>
<tr>
<td>( M[%] )</td>
<td>36.8</td>
<td>29.7</td>
</tr>
</tbody>
</table>

where \( R(T + 1) \) is the return of the stock invested at \( T \), and \( R(T + 1) \) is our return obtained by closing the position at \( T + 1 \). Then, \( \{ \cdot \} \) means a set, and its number, its average, and its sum are respectively denoted as \( |\{\cdot\}| \), \( \{\{\cdot\}\} \), and \( \sum\{\cdot\} \). In addition, we calculate the maximum drawdown rate \( M \), which is the maximum rate of the cumulative losses since the previous largest profit.

As a result, all strategies obtained positive returns. In particular, the strategies of SaN and BaN show better performances because these applied the nighttime’s abnormal price movements.

5. Statistical Significance Tests for Trading Strategy

Although each trading strategy is profitable by foreseeing the reverse movement to the abnormal price, we next confirm their validity by performing some statistical significance tests from two viewpoints: the trading position (short-sell or buy) and its trading timing.

5.1. Validity of trading positions

First, we investigate whether each trading strategy correctly decides the trading position of short-sell or buy. For this purpose, we randomly shuffle all of the original positions but keep each trading timing, and then we evaluate the trading performance using the same measures as in Sec. 4. For example, if we use \( R \) as a measure, we calculate \( R_r \) \( (r = 1 \sim 10,000) \) for 10,000 randomized strategies. Then, we compare them with \( R \) given by the original strategy in the statistical significance test. This method is a sort of nonparametric bootstrap test.

However, by randomly shuffling the trading positions, the number of each position decided by the randomized strategies becomes different from that for the original strategy. This difference gives either advantage or disadvantage to the randomized strategies. Therefore, we have to modify each investment performance given by the \( r \)th randomized
strategy. For example,
\[
\tilde{R}_r = \frac{1}{F}[R_r + \alpha],
\]
\[
W_r = W_r + \frac{1}{F} \sum_{i=1}^{N} \beta_i,
\]
\[
\beta_i = (B_{o,i} - B_{r,i}) \cdot \tilde{R}_i - (S_{o,i} - S_{r,i}) \cdot \tilde{R}_i,
\]
\[
\psi_r = \phi_r \cdot \frac{W_r}{1 - W_r},
\]
where \( \tilde{R}_i \) is the average return rate of the \( i \)th stock during the investment period. Then, \( B_{o,i} \) and \( S_{o,i} \) are the numbers of buying and selling the \( i \)th stock by the original strategy, and \( B_{r,i} \) and \( S_{r,i} \) are those by its \( r \)th randomized strategy. Similarly,
\[
\bar{R}_r = \sum_{i=1}^{N} \alpha_i
\]
\[
\tilde{\beta}_i = (B_{o,i} - B_{r,i}) \cdot \tilde{R}_i - (S_{o,i} - S_{r,i}) \cdot \tilde{R}_i,
\]
\[
\tilde{w}_i = \frac{|T[r_i(T+1) > 0]| - |T[r_i(T+1) < 0]|}{|T[r_i(T+1) \neq 0]|},
\]
where Eq. (19) is modified by using \( W_r \) given in Eq. (16), but \( \phi \) does not need any modification because it is calculated only by the average values in Eq. (10), and therefore the numbers of buy and sell positions are not essential. Then, it is so hard to modify \( M \) that the maximum draw down is not used in this test.

By comparing the trading performance given by the original strategy with those given by its randomized strategies, we can evaluate the \( p \)-value, that is, the percentage of random strategies that have better performance than the original strategy. If the \( p \)-value is less than 5\%, we can conclude that the original strategy is statistically valid. Table 4 shows each \( p \)-value in terms of several measures. We can confirm the validity of trading positions except in the strategy of SaD.

5.2. Validity of trading timings

Next, we investigate whether each strategy can decide trading timings correctly. Similarly to Sec. 5.1, we compose many randomized strategies, but in this section we only randomize the original trading timings. Therefore, the modifications by Eqs. (12)–(19) are unnecessary because the number of trading positions does not change from the original strategy. Then, we evaluate the \( p \)-value by comparing the original strategy with its randomized strategies. Table 5 shows each \( p \)-value. Similarly to Sec. 5.1, the strategies of SaN and BaN using the nighttime’s price movements show stronger validities in terms of not only the trading position but also its trading timing.

6. Conclusion

To detect abnormal stock prices, we proposed an indicator based on the autoencoder, which can learn the relationships among stocks and therefore can notify us when a stock price behaves abnormally. Then, by analyzing real stock data, we found that the following price movement is often biased right after the abnormal price. In particular, this tendency was clearly observed as the reaction to the nighttime’s falling price movement. The reason might be because of the asymmetric human mind and the closed market at nighttime. After that, we discussed trading strategies based on this anomaly and confirmed their validity by statistical significance tests in terms of trading positions and trading timings. All the given results are expected to be general because they were almost the same in the two major markets of Japan and USA.

This research was partially supported by JSPS KAKENHI Grant Number 16K00320.

References


An Alternative to Basic Log-likelihood for Bayesian Network Clustering
Rei Oshino† and Koujin Takeda‡
†Department of Intelligent Systems Engineering, Ibaraki University
4–12–1, Nakanarusesawashi Hitachi, 316-8511 Ibaraki, Japan
Email: 16nm912r@vc.ibaraki.ac.jp, koujin.takeda.kt@vc.ibaraki.ac.jp

Abstract—We study clustering problem of vertices on graph by Bayesian inference. In Bayesian framework of clustering, stochastic block model is a standard for construction of likelihood. Here we start with a variant of stochastic block model by Karrer and Newman for theoretical discussion. It is known that naive log-likelihood from their model does not always give natural clustering. We discuss how to modify it by adding correction term. By numerical experiment, we verify advantage of our method.

1. Introduction

Clustering is one of the main topics of unsupervised machine learning. In particular, clustering of vertices on graph, which we refer to as network clustering in this article (often termed as community detection generally), is also of great importance for extracting blockwise structure of real world network.

The idea of network clustering is not clearly defined mathematically in general. Its goal is to extract natural clusters for human sense. Toward this goal, several methods have been proposed and frequently used such as maximization of modularity[1] or spectral clustering[2].

Among them, we focus on Bayesian inference. In this framework, a probabilistic model of network under given network parameter is necessary, and stochastic block model (SBM)[3, 4] is a standard for the purpose. There are many variants of SBM in the definition of network probability, and we follow the one by Karrer and Newman[5] for theoretical discussion. The advantage of their SBM is its analytical simplicity: We can optimize some model parameters analytically and remove them from the model. As a result, remaining parameters are only the numbers of edges between clusters and the numbers of vertices in one cluster, and we maximize log-likelihood with respect to them to extract clusters. However it is reported that such naive approach sometimes yields unnatural clusters. One way to cope with this problem is the replacement of their SBM to another, while we must avoid overfitting due to overcomplex model with many additional model parameters.

We attempt another approach here: We make an appropriate choice of information criterion. In general Bayesian inference framework, several famous information criteria are frequently used, where correction term is added to log-likelihood. In our problem the same approach can be taken naturally. However, it should be noted that for network clustering there may be an appropriate information criterion among many possibilities.

In this article we propose a form of correction term for natural network clustering, which originates from intuitive discussion. Using log-likelihood with our correction term, we conduct numerical experiment of network clustering for real network data, and compare the result with naive method. We also discuss the relationship of our method with other works[6, 7, 8], where other information criteria are derived and proposed by theoretical argument.

2. Model

2.1. Naive SBM

The probability of naive SBM in [5] is given by

\[ P_{SBM}(G; \omega, g) := \prod_{i < j} \left( \frac{\omega_{g_i,g_j} A_{i,j}}{A_{i,j}!} \right)^{A_{i,j}} \exp\left(-\omega_{g_i,g_j}\right) \times \prod_{i} \left( \frac{1}{2} \frac{\omega_{g_i,g_i}}{A_{i,i}/2!} \right)^{A_{i,i}/2!} \exp\left(-\frac{1}{2} \omega_{g_i,g_i}\right). \] (1)

Roman subscripts \(i, j\) describe vertices on network. \(A_{i,j}\) is \((i,j)\)-element of network adjacency matrix. \(g_i\) is cluster index to which vertex \(i\) belongs. \(\omega_{g_i,g_j}\) is the expectation value of the number of edges between clusters \(g_i\) and \(g_j\), under the condition that the number of edges between vertex pair is independently Poisson distributed. The symbols \(G, \omega\) and \(g\) on l.h.s. denote a network sample, the set of \(\omega_{g_i,g_j}\), and the set of \(g_i(=\text{cluster assignment index on each vertex})\), respectively.

Naive Bayesian network clustering is based on the maximization of the log-likelihood with respect to \(\omega, g\) under given network sample \(G\).

\[ L_{SBM}(G; \omega, g) := \log P_{SBM}(G; \omega, g). \] (2)

After maximization with respect to \(\omega\) analytically, the maximum log-likelihood is expressed as

\[ L_{SBM}(G; g) := \max_{\omega} \left\{ \log P_{SBM}(G; \omega, g) \right\} \]
\[ = \sum_{i,j} m_{i,j} \log \frac{m_{i,j}}{n_{i} n_{j}}, \] (3)
where
\[ m_{i,j} := \sum_{\alpha} A_{i,j} \delta_{g_i,\alpha} \delta_{g_j,\beta}. \] (4)
is the number of edges between clusters indexed by Greek letters $\alpha, \beta$, and $n_\alpha$ is the number of vertices in cluster $\alpha$. Then, our task is to maximize log-likelihood in (3) for a given $G$ with respect to cluster assignment $g$, or equivalently the set of variables $\{m_{\alpha \beta}, n_\alpha\}$.

Naive SBM in [5] is suitable for theoretical analysis or argument, however it generates the network without natural clusters. Therefore this model must be corrected.

### 2.2. Degree-corrected SBM

Degree-corrected SBM is also proposed in [5] as a variant of naive SBM, where the probability of network $G$ is replaced by

$$P_{\text{DC-SBM}}(G|\theta, \omega, g) := \prod_{i<j} \frac{(\theta_{ij}\omega_{g(\theta, \omega, g)})^{k_{ij}}}{A_{ij}} \exp(-\theta_{ij}\omega_{g(\theta, \omega, g)}) \times \prod_{i} \frac{1}{2} \frac{(\omega_{g(\theta, \omega, g)})^{k_{ii}}/}{(A_{i}/2)!} \exp\left(-\frac{1}{2} \theta_{ii}\omega_{g(\theta, \omega, g)}\right).$$

(5)

where $\kappa_\alpha$ is the sum of all vertex degrees in cluster $\alpha$. Although we introduce novel parameter set $\theta$, the final expression of log-likelihood is as simple as the original naive model: $n_\alpha$ is just replaced by $\kappa_\alpha$. Therefore, degree-corrected SBM is regarded as the simplest variant of naive SBM.

With this modification, we can extract more natural clusters in comparison with naive SBM, however it still fails to extract clusters in some graphs.

### 3. Method

We start our discussion with log-likelihoods of naive SBM (3) or degree-corrected SBM (6). For appropriate information criteria, we add correction terms to these log-likelihoods as follows,

$$L_{\text{SBM}}^*(G|g) := L_{\text{SBM}}(G|g) + \sum_{\alpha} m_{\alpha} \log m_{\alpha} - \sum_{\alpha \neq \beta} m_{\alpha \beta},$$

(7)

$$L_{\text{DC-SBM}}^*(G|g) := L_{\text{DC-SBM}}(G|g) + \sum_{\alpha} m_{\alpha} \log m_{\alpha} - \sum_{\alpha \neq \beta} m_{\alpha \beta}. $$

(8)

Intuitively, the first correction term enhances the density of edges inside single cluster, and the second correction term penalizes the intertwining edges between clusters. By maximizing the log-likelihood with these correlations, extraction of more natural clusters is expected.

The advantage of such correction terms is that we do not need to introduce novel network parameters. We only use $m_{\alpha \beta}$ in correction terms, which has already been in the original log-likelihood. This is desirable for avoiding over-complex modeling. On the other hand, we should keep in mind that arbitrary multiplication factors can be introduced to correction terms. We set them unity in this article.

### 4. Numerical Experiment

#### 4.1. Clustering of artificial network

First, we prepare several artificial small-size networks, in which clusters are obvious. An example of artificial network is shown in Figure 1. We extract the clusters by maximizing $L_{\text{DC-SBM}}$ or $L_{\text{DC-SBM}}^*$, where the number of clusters is known in advance. For maximization, we conduct exhaustive search in this experiment.

The result is depicted in the same figure. Log-likelihood with correlation $L_{\text{DC-SBM}}^*$ yields natural clusters, whereas naive log-likelihood $L_{\text{DC-SBM}}$ gives unnatural ones. We also attempted other artificial networks having similar network/cluster structure, and found that $L_{\text{DC-SBM}}^*$ is also more successful. Hence we conclude that our method can cure the defect in the naive model.

In Figure 1, vertices in the same cluster are not mutually intertwined by many edges. By naive SBM, such networks will be generated only with very low probability, therefore we cannot extract natural clusters. By incorporating correction terms, we can avoid unnatural results.

#### 4.2. Clustering of small size network dataset

Next we apply our method to real network data of relatively small size (the number of vertices is $10^4 \sim 10^5$).
Table 1: Clustering result of Zachary’s “Karate club” dataset and “Dolphin” dataset. We search global maximum of $L_{DC-SBM}(G|g)$ or $L^*_{DC-SBM}(G|g)$ with 100 random initial conditions. We show the successful ratio of three cases: (A) Global maximum is found only for $L_{DC-SBM}(G|g)$. (B) Only for $L^*_{DC-SBM}(G|g)$. (C) For both.

We use Zachary’s “Karate club” dataset[9] (34 vertices), and “Dolphin” dataset[10] (62 vertices). There is the correct answer of clustering only for “Karate club” dataset, where two clusters exist. We extract the clusters on these networks by maximizing $L_{DC-SBM}$ or $L^*_{DC-SBM}$ and by assuming again the number of clusters is two and known, although “Dolphin” dataset is thought to have more communities (i.e. we focus on the largest two clusters).

For extraction, we use Kernighan-Lin algorithm[11]. In this algorithm we select the most relevant vertex to increase $L_{DC-SBM}(G|g)$ or $L^*_{DC-SBM}(G|g)$ by the change of its cluster assignment, then actually move it to another cluster. We repeat it until reaching local (or sometimes global) maximum of $L_{DC-SBM}(G|g)$ or $L^*_{DC-SBM}(G|g)$. We perform this algorithm under 100 random initial cluster assignments, and among 100 final resulting assignments we regard the largest $L_{DC-SBM}(G|g)$ or $L^*_{DC-SBM}(G|g)$ as global maximum, which appears under several initial conditions.

As a result, we finally reach the same cluster assignments both by $L_{DC-SBM}(G|g)$ and by $L^*_{DC-SBM}(G|g)$ for “Karate club” and “Dolphin” datasets. In addition, the final cluster assignments are natural. (For “Karate club”, only one vertex is misclassified.) However, the successful ratios of reaching global maximum from 100 initial assignments are different, as summarized in Table 1. This means that natural cluster assignment can be found by corrected log-likelihood $L^*_{DC-SBM}(G|g)$ more easily than the naïve one $L_{DC-SBM}(G|g)$.

### 4.3. Clustering of large size network dataset

Next we apply our method to real network data of relatively large size (the number of vertices is $10^2 \sim 10^3$). We use the datasets as follows:

- “football”[12] (114 vertices, 1224 edges)
- “euroroad”[13] (1174 vertices, 2834 edges)
- “netscience”[14] (1460 vertices, 5484 edges)

Here we assume the number of clusters (denoted by $K$) is unknown, which must be determined. For this purpose, we maximize the following quantities with respect to $K$:

- $L_{SBM}$ in (2)
- $L^*_{SBM}$ in (7)
- Factorized Information Criterion (FIC)[7, 8]: It is proposed for factorized asymptotic Bayesian inference and can be applied to network clustering by SBM for determination of optimal $K$. 

![Figure 2: Clustering result of large size network dataset. Top: "football", Middle: "euroroad", Bottom: "netscience".](image-url)
• FIC with our correction term like (7) and (8) (denoted by FIC in Figure 2)

In this experiment, we first fix $K$ and maximize these quantities. For maximization, we perform simulated annealing under a single random initialization of cluster assignment. Then we vary $K$ and search the maximum with respect to $K$ as well.

The result is depicted in Figure 2. The ratio of each quantity to $K = 10$ is shown on the vertical axis. It should be noted that the maximum of the original quantity is depicted as minimum, because all quantities are computed as negative.

For "football", $L_{\text{SBM}}$ and $L_{\text{SBM}}^*$ do not give the minimum. On the other hand, ICL and ICL* give the minimum around $K = 10$. The result of $K = 10$ seems quantitatively natural for the size of this dataset from the viewpoint of unsupervised learning, although we do not show the picture of clustering result here. For ICL*, the curvature of the graph around minimum seems larger than ICL, therefore this implies our method helps the determination of $K$ by an arbitrary optimization algorithm.

We cannot find a clear minimum for other two datasets, however, for "euroroad" there might be a minimum in the range of $10 < K < 20$, and the experiment with high precision will be desired. For "netscience", the experiment of higher $K$ will be necessary for finding a minimum.

5. Summary and discussion

We proposed an alternative to naïve log-likelihood for network clustering. By the result of numerical experiment, we verified our log-likelihood with correction yields more natural result of network clustering, or enables us to find natural clusters more easily.

Information criteria for network clustering are also proposed in preceding works, and FIC is one of them. In [8], the higher order correction to FIC under Laplace approximation is calculated under sparse network structure. Their resulting criterion is termed $F^3IC$ in their article, where the number of edges between clusters plays a significant role like our method. These criteria are obtained analytically, and we must discuss the relation between our method with them as a future work.

Acknowledgments

We are thankful to T. Kawamoto for his comment on the related work[8]. This work is supported by KAKENHI Nos. 24700007, 25120013 (KT).

References


A Construction Method for the Steiner Tree Problem using Betweenness Centrality

Misa Fujita†, Takayuki Kimura and Kenya Jin’no‡

†Graduate School of Electronics, Information and Media Engineering, Nippon Institute of Technology
4-1 Gakuendai, Miyashiro, Minami-Saitama, Saitama 345-8501, Japan
‡Department of Electrical and Electronic Engineering, Nippon Institute of Technology
4-1 Gakuendai, Miyashiro, Minami-Saitama, Saitama 345-8501, Japan
Email: 2158019@estu.nit.ac.jp, {tkimura, jinno}@nit.ac.jp

Abstract—Given an undirected weighted graph \( G = (V, E, c) \) and a set \( T \subseteq V \), where \( V \) is the set of vertices, \( E \) is the set of edges, \( c \) is the cost function, and \( T \) is the subset of terminal vertices, the Steiner tree is a subgraph that doesn’t have a loop and connects all terminals. The Steiner tree problem in graphs is to find the minimum cost Steiner tree. The Steiner tree problem is one of the \( NP \)-complete combinatorial optimization problems. Thus, approximate methods are usually employed for constructing the Steiner tree. In this study, the KMB algorithm, which is an efficient construction method for the Steiner tree problem, is enhanced by considering betweenness centrality. Results of numerical simulations indicate that our improved KMB algorithm shows good performance for the benchmark Steiner tree problems.

1. Introduction

Given an undirected weighted graph \( G = (V, E, c) \) and a set of \( T \subseteq V \), where \( V \) is the set of vertices, \( E \) is the set of edges, \( c \) is a non-negative cost function, and \( T \) is a subset of terminal vertices, a subgraph that doesn’t have a loop and connects all terminals is called a Steiner tree. A cost of the Steiner tree is the sum of the edge costs included in the tree. The Steiner tree problem in graphs is to find the minimum cost Steiner tree. Figure 1 shows examples of the Steiner tree. In Fig. 1, circles express vertices and lines express edges. Black circles express terminal vertices and black lines express the edges in the Steiner tree. A network has many Steiner trees. Among them, Fig. 1(c) is the minimum cost Steiner tree for the given network.

![Figure 1: Examples of the Steiner tree](image)

Figure 1: Examples of the Steiner tree

The Steiner tree problem is applied to various real-world problems, such as VLSI routing[1], wirelength estimation[2], and network routing[3]. Because the Steiner tree problem is one of the \( NP \)-complete combinatorial optimization problems[4], the approximation method is usually used to obtain the Steiner tree. Most popular method to obtain the near-optimum Steiner tree is the KMB algorithm[5]. The KMB algorithm constructs the Steiner tree based on the shortest path between terminals and the minimum spanning tree of all the terminals. The calculation cost of the KMB algorithm becomes \( O(|V|^2) \) where \( |V| \) is the number of vertices. The KMB algorithm can obtain small cost Steiner tree in short time. However, if there are some shortest paths between the terminals, an obtained cost of the Steiner tree changes depending on the selected paths.

From this viewpoint, we propose a construction method of the Steiner tree using the betweenness centrality in this study. The cost of the Steiner tree obtained by the KMB algorithm using given edge cost is compared with the one using the betweenness centrality of nodes and edges. Results of numerical simulations indicate that our propose method can obtain the small cost Steiner tree.

2. Steiner tree problem

If the number of terminals \( |T| \) becomes 2, the Steiner tree problem changes the shortest path problem. In addition, in the case by \( |T| = |V| \), the Steiner tree problem becomes minimum spanning tree problem. On the other hand, in the case by \( 2 < |T| < |V| \), finding minimum cost tree becomes the Steiner tree problem.

To mathematically model this problem, we first define a decision variable, \( x(e_i) \), that is defined as follows:

\[
x(e_i) = \begin{cases} 
  1 & (e_i \in E_{ST}), \\
  0 & \text{(otherwise)},
\end{cases}
\]

(1)

where \( e_i \) is the \( i \)-th edge in the network, \( E_{ST} \) is the set of edges included in a Steiner tree. By using the decision variables, an objective function for the Steiner tree problem in graphs is then defined as follows:

\[
\min \sum_{i=1}^{|E|} c(e_i)x(e_i),
\]

(2)
where $|E|$ is the total number of edges, $c(e_i)$ is a cost of the $i$th edge. If the $i$th edge, $e_i$, is included in the Steiner tree, a corresponding decision variable takes 1, and the cost of the edge $e_i$ is added to the total cost of the Steiner tree. In other words, the Steiner tree problem is to find the best combination of $x(e_i)$.

3. KMB algorithm

In this study, we use the KMB algorithm [5] to construct the Steiner tree. The KMB algorithm is one of the efficient approximation algorithms for the Steiner tree problems. To obtain the Steiner tree using the KMB algorithm, we first construct a complete graph $G_1 = (T_1, E_1, c_1)$ from $G$ and $T$ (Fig. 2(b)). In the complete graph $G_1$, all the vertices are terminals. The cost of the edges in $G_1$ corresponds to the cost of the shortest path between these terminals. Next, we construct a minimum spanning tree $T_1$ from the complete graph $G_1$ (Fig. 2(c)). Finally, edges used in the $T_1$ are replaced by the shortest paths connecting to the corresponding vertices in original graph (Fig. 2(d)). An example Steiner tree using the KMB algorithm is shown in Fig. 2.

![Figure 2: An example Steiner tree using the KMB algorithm](image)

4. KMB algorithm using betweenness centrality

In the KMB algorithm, if there are some shortest paths that have the same cost connecting to two vertices, the cost of the obtained Steiner tree changes depending on the selected paths (Fig. 3). The cost of the obtained Steiner tree might be longer if an undesirable path is selected. Thus, it is important to select better shortest paths which various shortest paths go through as much as possible.

From this viewpoint, we use the network centrality for constructing the shorter Steiner tree. Although various types of network centrality have already been proposed, we employ the betweenness centrality [6, 7, 8] in this study.

The betweenness centrality expresses how many times a vertex appears on every shortest path in the graph. We modify the cost of link to combine the original cost and the betweenness cost.

The betweenness centrality of a vertex $v$ is defined as follows:

$$bc(v) = \sum_{s=1}^{[V]-1} \sum_{g=1}^{[V]-g} \frac{P_{s,g}(v)}{(V-1)(V-2)/2},$$

where $s$ is a start vertex and $g$ is a goal vertex of the shortest path, $[V]$ is the number of vertices, $P_{s,g}(v)$ is the number of shortest paths between $s$ and $g$, $P_{s,g}(v)$ is the number of shortest paths between $s$ and $g$ that goes through the vertex $v$.

If a vertex has high betweenness centrality, the vertex frequently lies on the shortest path. By selecting the path includes the high betweenness vertices, the cost of the obtained Steiner tree may become small.

To use the betweenness centrality of vertices to edges, the betweenness cost of the edge $(v_i, v_j)$, $eb(v_i, v_j)$, is defined as follows:

$$eb(v_i, v_j) = \frac{bc(v_i) + bc(v_j)}{2},$$

where $v_i$ is the $i$th vertex, $v_j$ is the $j$th vertex, $bc(v)$ is the betweenness centrality of the vertex $v$.

We then define a new cost of edge by considering the given edge cost and the betweenness cost as follows:

$$c_{new}(e_i) = ac(e_i) + (1 - a) \frac{1}{bc(e_i)},$$

where $c(e_i)$ is the normalized given edge cost of the $i$th edge, $bc(e_i)$ is the normalized betweenness cost of the $i$th edge, $a$ is a controlling parameter which determines the priority between the given edge cost and the betweenness cost. Although the betweenness centrality is usually calculated by using the all the shortest paths in the graph, vertices placed at the outside from every terminal are not necessary to include the Steiner tree. We then limit the area of the graph for calculating the betweenness centrality in this study. Figure 4 illustrates the given network and the vertices for calculating the betweenness centrality (inside the
Table 1: Result of numerical simulations

<table>
<thead>
<tr>
<th>Name</th>
<th>Opt</th>
<th>Min</th>
<th>Max</th>
<th>Mid</th>
<th>α</th>
<th>Min</th>
<th>Max</th>
<th>Mid</th>
<th>KMB algorithm</th>
<th>KMB algorithm using Betweenness centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>344</td>
<td>374</td>
<td>387</td>
<td>374</td>
<td>0.6</td>
<td>364</td>
<td>364</td>
<td>364</td>
<td></td>
</tr>
<tr>
<td>dmxa0296</td>
<td>1017</td>
<td>1036</td>
<td>1046</td>
<td>1046</td>
<td>0.1</td>
<td>1031</td>
<td>1031</td>
<td>1031</td>
<td>1031</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>914</td>
<td>991</td>
<td>1009</td>
<td>1004</td>
<td>0.1</td>
<td>958</td>
<td>958</td>
<td>958</td>
<td></td>
</tr>
<tr>
<td>dmxa0454</td>
<td>275</td>
<td>297</td>
<td>297</td>
<td>297</td>
<td>0.2</td>
<td>277</td>
<td>277</td>
<td>277</td>
<td>277</td>
<td></td>
</tr>
<tr>
<td>dmxa0628</td>
<td>506</td>
<td>537</td>
<td>541</td>
<td>537</td>
<td>0.2</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td>537</td>
<td></td>
</tr>
<tr>
<td>dmxa0734</td>
<td>594</td>
<td>637</td>
<td>657</td>
<td>644</td>
<td>0.1</td>
<td>632</td>
<td>632</td>
<td>632</td>
<td>632</td>
<td></td>
</tr>
<tr>
<td>dmxa0848</td>
<td>580</td>
<td>639</td>
<td>660</td>
<td>660</td>
<td>0.3</td>
<td>645</td>
<td>645</td>
<td>645</td>
<td>645</td>
<td></td>
</tr>
<tr>
<td>dmxa0903</td>
<td>1488</td>
<td>1538</td>
<td>1561</td>
<td>1551</td>
<td>0.1</td>
<td>1546</td>
<td>1546</td>
<td>1546</td>
<td>1546</td>
<td></td>
</tr>
<tr>
<td>dmxa1010</td>
<td>454</td>
<td>473</td>
<td>498</td>
<td>483</td>
<td>0.2</td>
<td>473</td>
<td>473</td>
<td>473</td>
<td>473</td>
<td></td>
</tr>
<tr>
<td>dmxa1109</td>
<td>750</td>
<td>821</td>
<td>834</td>
<td>821</td>
<td>0.1</td>
<td>769</td>
<td>769</td>
<td>769</td>
<td>769</td>
<td></td>
</tr>
<tr>
<td>dmxa1200</td>
<td>311</td>
<td>329</td>
<td>334</td>
<td>334</td>
<td>0.1</td>
<td>316</td>
<td>316</td>
<td>316</td>
<td>316</td>
<td></td>
</tr>
<tr>
<td>dmxa1304</td>
<td>508</td>
<td>513</td>
<td>533</td>
<td>523</td>
<td>0.3</td>
<td>528</td>
<td>528</td>
<td>528</td>
<td>528</td>
<td></td>
</tr>
<tr>
<td>dmxa1516</td>
<td>780</td>
<td>825</td>
<td>840</td>
<td>830</td>
<td>0.2</td>
<td>794</td>
<td>794</td>
<td>794</td>
<td>794</td>
<td></td>
</tr>
<tr>
<td>dmxa1721</td>
<td>1365</td>
<td>1484</td>
<td>1540</td>
<td>1530</td>
<td>0.4</td>
<td>1479</td>
<td>1479</td>
<td>1479</td>
<td>1479</td>
<td></td>
</tr>
</tbody>
</table>

In Fig. 4, black circles express the terminal vertices. Here, $V_{lim}$ is the set of vertices in the limited area, and $E_{lim}$ is the set of edges connected to two vertices in $V_{lim}$. We call the betweenness centrality calculated using the limited graph $G_{lim} = (V_{lim}, E_{lim}, c)$ as a limited betweenness centrality.

5. Numerical experiments

To compare the performance of the conventional KMB algorithm with our proposed method, we used the benchmark problems provided in SteinLib[9]. The value of the controlling parameter $\alpha$ in our proposed method is set to the optimal value by the preliminary experiments. A source vertex of the Steiner tree is randomly selected among the terminals.

Table 1 shows the cost of the obtained Steiner trees by the conventional KMB algorithm and the proposed method for the test sets DMXA in SteinLib[9]. From the results, our proposed method obtained the shorter Steiner tree than the conventional KMB algorithm. In addition, the method using betweenness centrality and that by limited betweenness centrality show similar performance. However, the limited betweenness centrality has less calculation cost than the original betweenness centrality. Thus, we can say that the method using the limited betweenness centrality is better constructing method than that by the original betweenness centrality.

Figure 5 shows the performance of our propose method for various $\alpha$ cases. In Fig. 5, the constructing method by the limited betweenness centrality shows good performance almost the $\alpha$ cases. This is because the edges in the outside of limited area are removed.

Figure 6 shows the topologies of Steiner tree obtained by the conventional method and our propose method. Black circles denote the terminals and black lines denote edges in the Steiner tree. In Fig. 6, these Steiner trees have different topologies.
6. Conclusions

In this study, we try to change the edge cost by using the betweenness centrality for constructing the small cost Steiner tree. We evaluated the performance of our propose method by the testsets in the SteinLib. From the results of numerical simulations, the propose method obtains the small cost Steiner tree stably in comparison to the conventional KMB algorithm. The conventional KMB algorithm often finds larger cost Steiner tree if there are some shortest paths between the nodes. By selecting the shortest path which has high betweenness centrality, the cost of the Steiner tree becomes small. In addition, although the calculation cost of the betweenness centrality becomes large as the size of network increases, we can solve this undesirable problem by using the limited betweenness centrality.

The performance of the KMB algorithm can be enhanced by selecting the shortest paths between the terminals which various shortest paths commonly use. Because our propose method changed the costs of edges to improve the performance of the KMB algorithm, obtaining the smallest cost path by using the original edge cost is impossible. If we employ the edge which has original edge cost and high betweenness centrality, the cost of the obtained Steiner tree might become small. In the future work, we try to evaluate our method by the network whose links have original cost and high betweenness centrality to construct the small cost Steiner tree.

The research of T.K. was partially supported by a Grant-in-Aid for Young Scientists(B) from JSPS (No. 16K21327).

References


On constructing networks from multivariate nonlinear time series

Tomomichi Nakamura, 1,∗ Toshihiro Tanizawa, 2,† and Michael Small 3,4,‡

1 Graduate School of Simulation Studies, University of Hyogo
7-1-28 Minatojima-minamimachi, Chuo-ku, Kobe, Hyogo 650-0047, Japan
2 Kochi National College of Technology
200-1 Monobe-Otsu, Nankoku, Kochi 783-8508 Japan
3 School of Mathematics and Statistics, The University of Western Australia,
Crawley, WA 6009, Australia
4 Mineral Resources, CSIRO, Kensington, WA 6151, Australia
Email: ∗ tomo@sim.u-hyogo.ac.jp, † tanizawa@ee.kochi-ct.ac.jp, ‡ michael.small@uwa.edu.au

Abstract—We introduce a method for constructing networks from multivariate nonlinear time series from a deterministic dynamical systems perspective. The method can be applied even when the data exhibit no obvious qualitative similarity: a situation in which the naive method utilizing the cross correlation function directly cannot correctly identify connectivity. The method is demonstrated for numerical data sets generated by known systems and applied to several experimental time series.

1. Introduction

To understand the nature of ongoing interaction in real-world complex systems it is first necessary to deduce the interconnection between the components of the system (or underlying system) under study [1]. Elements in the system interact with each other. Once the connectivity has been determined the effect of that connectivity is frequently studied using the concept of complex networks [2].

There are also approaches for constructing networks for multivariate time series [3, 4]. In these approaches each time series is considered as a basic node of a network. Nodes are connected if the dynamics of the corresponding scalar time series are sufficiently similar. The naive (and usual) way to measure “similarity” between two signals is with the cross correlation function with a fixed threshold [3, 4]. While this naive approach is expeditious, it is also flawed when one is looking at nonlinear (possibly temporally delayed) interaction in complex systems (in other words, two signals are not similar). We describe the naive approach in detail in Sec. 2.

The most important thing to investigate the relationship between two signals is not similarity but correlation structures from the viewpoint of a deterministic dynamical system. Recently a method to construct such a network for multivariate nonlinear time series has been proposed based on this perspective [5]. To verify the intrinsic (essential) connection between two data sets, the previously proposed small shuffle surrogate (SSS) method is applied in the method, which can investigate correlation structures irrespective of whether the structures are linear or nonlinear [6]. That is, the method constructs networks for multivariate time series, even if there are nonlinearities in these time series.

2. The naive approach to network construction

The most extensively used method to construct networks for multivariate time series can be reduced to the following three steps [3, 4].

1. Each time series is considered as a basic node of a network.
2. To investigate the relationship among multivariate time series, the cross correlation between each pair of time series (i.e. two time series) taken from the whole multivariate time series is estimated.
3. The pair of nodes corresponding to the chosen two time series is connected with an undirected edge when the value of the cross correlation is larger than an appropriately chosen threshold.

We refer to this method as “the naive method.” The basic idea behind the naive method is as follows. When signals are similar, we expect that there may be some sort of relationship between the corresponding nodes, and hence the pair is considered to be connected with an undirected link. On the other hand, there are cases where time series are not similar enough. In this case, as we may have the impression that these are independent or have no relationship, we do not connect them. This approach relies on selecting an appropriate threshold. Although the naive method has been proved to be effective in various cases [3, 4], the range of applicability might be restrictive because “no similarity” is not equivalent to “no correlation” and “no relationship” [6]. Furthermore, there is a possibility that “similarity” might not be equivalent to “relationship.” That is, the naive method cannot deal with data appropriately especially when there are nonlinearities. In the next section, we describe an approach to reduce this problem.
3. A different approach to construct networks

The approach of the proposed method is basically the same as the naive method described in Sec. 2. The difference is the way of verifying the connection between two data sets. As mentioned above, only the cross correlation function is used in the naive method. To determine whether two nodes should be connected statistically and to make the result rigorous, we apply the small shuffle surrogate (SSS) method [6], because the SSS method has broad applicability\(^1\) and can examine whether there are correlation structures\(^2\).

3.1. The small-shuffle surrogate method

To investigate whether temporal correlations in time series data are absent or if the data are independently distributed random variables, the SSS method is often used [6]. The SSS method destroys local structures or correlations in irregular fluctuations (short-term variabilities) and preserves the global behaviours by shuffling the data index on a small (local) scale.

SSS data are generated as follows. Let the original data be \(x(t)\), let \(i(t)\) be the index of \(x(t)\) [that is, \(i(t) = t\), and so \(x[i(t)] = x(t)\)], let \(g(t)\) be Gaussian random numbers and \(x(t)\) will be the surrogate data.

(i) Obtain \(i'(t) = i(t) + Ag(t)\), where \(A\) is an amplitude.

(ii) Sort \(i'(t)\) by the rank-order and let the index of \(i'(t)\) be \(i'(t)'\).

(iii) Obtain the surrogate data \(s(t) = x[i'(t)']\).

It has been found that choosing \(A = 1.0\) is adequate for nearly all purposes [6] — although this parameter choice remains heuristic. Further details of the method and the mechanism are provided in [6]. When we apply the SSS method to multivariate data, the null hypothesis (NH) is that there is no short-term correlation structure between the data or that the irregular fluctuations are independent [6].

3.2. When to reject a null hypothesis

Discriminating statistics are necessary for surrogate data hypothesis testing. The SSS method changes the flow of information in the data. It is preferable to use discriminating statistics which can accurately reflect features of the surrogate method. Hence, we choose to use the cross correlation (CC) function and the average mutual information (AMI) as discriminating statistics. These statistics can determine, on average, how much one learns about one signal by observing the other [7].

After the calculation of these statistics, we need to determine whether a null hypothesis (NH) should be rejected. We employ Monte Carlo hypothesis testing and determine whether the estimated statistics of the original data fall within or outside the statistical distribution of the surrogate data [8]. When the statistics fall within the distributions of the surrogate data, we conclude that the hypothesis may not be rejected. In this paper, we generate 99 SSS data and hence the non-parametric significance level is between 0.01 and 0.02 for a one-sided test with two non-independent statistics\(^3\).

4. Numerical Example

We demonstrate the application of our algorithm to one simulated multivariate time series data set, and confirm our theoretical arguments with the several example. For comparison we also apply the naive method to the data sets. In this case, we use \(A = 1.0\) for generating SSS data, generate 99 SSS data, and the data is 1000 points with Gaussian observational noise with the mean zero and the standard deviation 0.01.

4.1. Data from a nonlinear system

To investigate whether the proposed method works even if there is nonlinearity, we use the system which consists of four dynamical variables, \(x_1(t), x_2(t), x_3(t),\) and \(x_4(t)\), and the models are described by the following expressions:

\[
\begin{align*}
x_1(t) &= 1.3 + 0.2 x_1(t - 1) - 0.1 x_1(t - 3) + 0.1 x_2(t - 4) x_4(t - 7) + e_1(t), \\
x_2(t) &= 2.0 + 0.6 x_2(t - 1) - 0.2 x_2(t - 6) + e_2(t), \\
x_3(t) &= h[2.2 + 0.2 x_1(t - 2) + 0.3 x_4(t - 9) + e_3(t)], \\
x_4(t) &= 1.3 + 0.2 x_1(t - 2) + 0.5 x_4(t - 1) - 0.3 x_3(t - 3) + e_4(t),
\end{align*}
\]

where \(e_i(t) (i = 1, 2, 3, 4)\) are dynamic noise, independent and identically distributed Gaussian random variables with mean zero and standard deviation 1.0. The function \(h(x)\) is a static monotonic nonlinear function [9],

\[
h(x) = \frac{5.0}{1 + e^{-0.0001 x} + e^{-0.0001 x} + a e^{-0.0001 x} + b e^{-0.0001 x}},
\]

where \(\rho = 3, a = -2.0\) and \(b = 10.0\). The behaviours of the four time series generated by these models are shown in Fig. 1. The behaviours show irregular fluctuations and it is difficult to know the relationship among the data by visual inspection.

\(^1\)The SSS method can investigate whether there are correlation structures in short-term variabilities among data, irrespective of whether data have similar or different long-term trends.

\(^2\)The term “correlation structures” we use means any structures, irrespective of whether the structures are linear or nonlinear.

\(^3\)The significance level of each test is 0.01. If two statistics are identical (dependent), the significance level for the proposed test is 0.01. If the statistics are independent, the significance level for the test is given by \(1.0 - 0.99 \times 0.99 = 0.0199\). Hence, the reality should be somewhere in-between [6].
Figure 1: Time series data generated by the nonlinear system, Eqs. (1)–(4). We use the data to construct the network.

In this paper, we distinguish between “component” and “variable” as different technical terms. We use the term “component” to represent $x_i$ and the term “variable” when it takes a particular value $x_i(t-i)$. We treat the components as the nodes of the network. That is, Eq. (1) has three components ($x_1$, $x_2$ and $x_4$) and four variables, $x_1(t-1)$, $x_1(t-3)$, $x_2(t-4)$, and $x_4(t-7)$. As shown in Eqs. (1)–(4), each dynamical variable at time $t$ is determined by various other dynamical variables. We consider the connectivity of the linear system, Eqs. (1)–(4). Eq. (1) shows that the component $x_1$ is influenced by three components, $x_1$, $x_2$ and $x_4$. That is, other components which connect the component $x_1$ are $x_2$ and $x_4$. Similarly, as Eq. (2) shows that $x_2$ is driven by only $x_2$, there is no connection with $x_2$. As Eq. (3) shows that $x_3$ is driven by $x_1$ and $x_4$, $x_1$ and $x_4$ connect $x_3$. As Eq. (4) shows that $x_4$ is driven by $x_1$ and $x_4$, $x_1$ connects $x_4$. Based on this, the connectivity expressions of the nonlinear system become

$$x_1 = f_1(x_2, x_4),$$  \(6\)
$$x_2 = 0,$$  \(7\)
$$x_3 = f_3(x_1, x_4),$$  \(8\)
$$x_4 = f_4(x_1),$$  \(9\)

where $f_i$ stands for the function representing connectivity of the $i$-th component, $x_i$, and zero means that there is no connection. The network structure constructed based on this idea is shown in Fig. 2(a).

We estimate the cross correlation function to apply to the naive method. All the values are shown in Table 1. We need to determine the fixed threshold value to decide whether a link is present between two components. If we set the value 0.5, as shown in Table 1, we cannot connect any link between nodes. The network structure constructed by the naive method is shown in Fig. 2(b), and Fig. 2(b) shows that there is no link among any node on this network. However, we note that as Eqs. (1)–(4) show, there are correlation structures among the components. This result clearly indicates that only the application of the cross correlation function is not effective.

We apply the SSS method to the data of all possible pairs to verify the connection between two data sets. Figure 3 shows the result. This result indicates that we can discriminate correctly whether there are correlation structures between two signals. Also, other data sets are discriminated correctly. Based on this we can construct the same network as shown in Fig. 2(a).

![Figure 2](image1.png)

**Figure 2**: (Colour online) The linkage of network: (a) the connectivity of Eqs. (1)–(4). The same network is obtained when the proposed method is applied to the data shown in Fig. 1. (b) the network when we apply the naive method to the data. As shown in this figure, there is no link among the nodes.

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>1.0000</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.3413 (-4)</td>
<td>1.0000</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.3337 (2)</td>
<td>0.0688 (6)</td>
<td>1.0000</td>
<td>—</td>
</tr>
<tr>
<td>$x_4$</td>
<td>0.4113 (-7)</td>
<td>0.0725 (8)</td>
<td>0.3906 (-9)</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Table 1**: The largest absolute values of the cross correlation function of all possible pairs between the time lag $-10$ and $+10$, where the number in parentheses is the time lag when the cross correlation function has the largest absolute value. The data are generated by the nonlinear system, Eqs. (1)–(4), and the values are estimated using 1000 data points.

![Figure 3](image2.png)

**Figure 3**: (Colour online) The result of nonlinear system, Eqs. (1)–(4). A plot of (a) and (c) the cross correlation function (CC), and (b) and (d) the average mutual information (AMI), (a) and (b) are result of $x_1$ and $x_2$, and (c) and (d) are result of $x_2$ and $x_3$. The solid line is the original data and the dotted lines are the SSS data.
5. Application

Based on the results of the computational studies, we apply the proposed method to one experimental system: hourly meteorological time series data set in Kobe, Japan. The meteorological time series data set are five different time series: the atmospheric pressure, the temperature, the dew-point temperature, the vapour pressure and the humidity taken hourly in Kobe, Japan from 1 January to 28 February in 2013\(^4\). As shown in Fig. 4, each of them shows irregular fluctuations. We use 1416 data points for a meteorological time series data set. In all cases we use \(A = 1.0\), generate 99 SSS data and estimate the CC function and the AMI between the time lag \(-10\) and \(+10\).

Figure 5 shows networks constructed by the naive method and the proposed method. There are interesting differences between them. The CC functions between the temperature and the dew-point temperature and between the temperature and the vapour pressure are larger than 0.5. Hence, these are connected as shown in Fig 5(a). However, as the CC and AMI of the original data fall within the distributions of SSS data, these are not connected as shown in Fig 5(b). This might indicate that “similarity” is not equivalent to “relationship.”

6. Conclusion

We have introduced an algorithm to construct networks for multivariate time series using the SSS method. The network constructed by the method indicates the intrinsic connectivity of the elements included in the system.

Acknowledgments

T.N would like to acknowledge the partial support of a Grant-in-Aid for Scientific Research (C) (No. 25282094) from the Japan Society for the Promotion of Science (JSPS), and T.T also would like to acknowledge the support of a Grant-in-Aid for Scientific Research (C) (No. 24540419) from JSPS. M.S is funded by the Australian Research Council via a Future Fellowship (FT110100896) and Discovery Project (DP140100203).

References


\(^4\)The data set can be obtained from Japan Meteorological Agency, http://www.jma.go.jp/jma/indexe.html
A new approach to non-uniformly sampled time series using ordinal partitions

Thomas Stemler†‡⋆, Konstantinos Sakellariou‡, Michael McCullough‡ and Ben Luo‡

†Complex Data Modelling Group, The University of Western Australia 35 Stirling Hwy, Crawley WA, 6009, Australia
‡School of Mathematics and Statistics, The University of Western Australia 35 Stirling Hwy, Crawley WA, 6009, Australia
⋆ Postam Institut für Klimafolgenabschätzung (PIK), RD IV Transdisciplinary Concepts and Methods Potsdam, Germany

Email: thomas.stemler@uwa.edu.au, konstantinos.sakellariou@research.uwa.edu.au

Abstract—The ordinal patterns symbolic methodology may be used to discriminate between deterministic and stochastic time series. Deterministic systems are characterised by forbidden patterns which would not occur even in infinitely long time series, whereas random systems may be exhibit any possible pattern. In this study, we examine the effects of highly irregularly-sampled time series, such those encountered in paleoclimate measurements, on the estimator of the proportion of forbidden patterns manifested by synthetic time series. We investigate three different sampling regimes and observe a high sensitivity in the results. The degree of asymmetry of the sampling distribution is shown to be a useful heuristic indicator for the reliability of the forbidden patterns estimator and the potential for detecting determinism. Exponentially distributed lead to gross underestimation of the total number of forbidden patterns. On the other hand, sampling in the presence of large chronological gaps can lead to relatively accurate estimates as long as the time series contains sufficiently many densely-sampled areas as well.

1. Introduction

Detecting a deterministic component in noisy data is an important problem in nonlinear time series analysis. The assumption of determinism underlies a large class of techniques which focus on the theory of dynamical systems. Symbolic dynamics tools have recently shown potential towards this goal. Ordinal patterns, in particular, comprise symbols obtained by a segmentation of a time series into elements of equal length. Patterns which cannot occur for a specified system are termed forbidden. Analysing the statistical properties of the resulting sequence can shed light on the underlying dynamics. Deterministic time series are thought to always be characterised by forbidden patterns, in contrast to random systems whereby any possible pattern may be realised. Thus, the relative proportion of forbidden patterns can be used to detect determinism. We investigate the effects of highly irregular sampling, such as from paleoclimate or geological data, on the reliability of this statistic as estimated from time series data. Real-world data sets often appear in the form of non-uniformly sampled time series. This may be due to device failure, weather conditions, human error, the nature of the system (e.g. financial transactions data) or the measurement method (e.g. geological data) and other causes. For this study we are motivated by geoscientific and paleoclimatic time series which are characterised by missing entries and large chronological gaps. Although there exist several types of irregular sampling, which can vary from rather mildly to highly unevenly spaced data, the majority of established techniques in time series analysis assume regular sampling. Consequently, there is an increasing need to extend the applicability of existing techniques and create more sophisticated tools to reliably analyse irregularly sampled time series. For linear systems, the Lomb-Scargle periodogram (also known as Vanicek’s least-square method) is an example towards this direction. For nonlinear systems, there exist a few notable (but very recent) examples such as the similarity estimators method (K. Rehfeld and J. Kurths, Climate of the Past 10, 107 (2014)), the distance metric for marked point processes by Suzuki, Hirata, and Aihara (S. Suzuki, Y. Hirata, and K. Aihara, Int. J. Bifurcation Chaos 20, 3699 (2010)) and the transformation-cost time series by Eroglu et al. (D. Eroglu, F. McRobie, I. Ozken, T. Stemler, K.-H. Wyrwoll, S. F. M. Breitenbach, N. Marwan, and J. Kurths, Nature Communications 7, 12929 (2016)).

In this paper, we further extend upon previous results by exploring cases of more severe irregular sampling, far removed from the regular sampling grid archetype. The types of unevenly sampled synthetic data we examine herein are more characteristic of geoscientific measurements such as geological or paleoclimatic data, and pose a formidable challenge for the ensuing time series analysis under any methodological approach. The question we ask is about the reliability of the estimator of forbidden patterns as a criterion for the existence of determinism in highly irregularly sampled time series. The sampling schemes we examine here include a Poisson sampling process (exponentially distributed in intervals), Pareto sampling (example of a power-law distribution) and F-distributed sampling periods for various levels of skewness.
String Entropy as a Measure of Complexity in Chaotic Time Series

Takaya Miyano and Kenichiro Cho
Department of Mechanical Engineering, Ritsumeikan University
1–1–1 Noji-higashi, Kusatsu, Shiga 525-8577, Japan
Email: tmiyano@se,ritsumeik.ac.jp, kcho@fc.ritsumeik.ac.jp

Abstract—Chaotic time series numerically generated by deterministic nonlinear dynamics are often used as pseudorandom numbers for cryptography. Here, we show a method for time series analysis to characterize complexities in chaotic dynamics in terms of string entropy, as a class of information entropy, estimated from the relative frequencies of binary-coded characters transformed from a chaotic time series. We apply our method to various chaotic time series and discuss their performance as pseudorandom numbers for chaos-based cryptography.

1. Introduction

Chaotic dynamics have been often applied to cryptography, where the chaotic time series numerically generated by the dynamics were used as pseudorandom numbers to encrypt plaintexts and message signals [1]–[12]. Recently, we have derived the augmented Lorenz (AL) equations as a nondimensionalized dynamical model for turbulent thermal convection with a high Rayleigh number exceeding $10^6$ from the Newtonian equations of motion of a chaotic gas turbine [13]. The AL equations have been applied to a one-time pad chaotic cryptography, where the chaotic time series generated by the AL equations were used as pseudorandom numbers to mask a speech signal and a plaintext [14].

The randomness of pseudorandom numbers is of critical importance when using them in cryptosystems to encrypt messages. A low degree of randomness will facilitate code breaking by cryptanalysts. In fact, the statistical tests for the randomness of pseudorandom numbers to be used in cryptosystems, i.e., NIST 800-22, is published by the National Institute of Standards and Technology (NIST) [15], which is widely recognized as the standard protocol to assess the randomness.

We have recently proposed an information-theoretical method, referred to as the string entropy method [16], to evaluate the degree of complexity in a chaotic time series. Although the string entropy test is essentially the same as the entropy test included in NIST 800-22, it appears to also be effective to characterize the dynamical nature of chaotic dynamics. In this paper, we apply our method to the chaotic time series numerically generated by various chaotic dynamics and show how it is effective to characterize the chaotic nature.

2. String Entropy

Given a time series $\{x_i\}_{i=1}^N$ with a sampling time interval of $T$, the string entropy $S$ is defined as follows. First, $x_i$ is transformed into binary digits $b_i$ with the threshold crossing as

$$b_i = 0 \text{ if } x_i < x_c,$$

$$b_i = 1 \text{ otherwise,}$$

where $x_c$ is an appropriately chosen threshold around which $b_i$ is distributed with equal probability. We partition the binary series $\{b_i\}_{i=1}^N$ into a sequence of $Q$ segments consisting of $D$ binary digits, where $N = DQ$. Each segment is mapped to an “alphabet” binary-coded in $D$ bits, where each alphabet symbolizes the time evolution represented by $D$ successive realizations of the dynamical system. Thus, we obtain a string of $Q$ alphabets as random realizations of $2^D$ possible alphabets denoted as $\{a_n\}_{n=1}^{2^D}$. In the case of $D = 7$, we have $2^7 = 128$ alphabets \{0000000, 0000001, \ldots, 1111111\} corresponding to \{0’, 1’ \ldots, 127’\} in the decimal expression. In this case, the total number of the alphabets is equal to that of the ASCII codes. With the frequency of appearance for each alphabet $a_n (n = 1, \ldots, 2^D)$, we estimate a histogram, from which the probability density function $p(a_n)$ is calculated. The histogram represents the statistical distribution of the coarse-grained time evolution. The string entropy $S$ is defined and calculated as

$$S = - \sum_{n=1}^{2^D} p(a_n) \log_2 p(a_n).$$

$S$ takes the maximum value $S_{\text{max}} = D$ [bits] if and only if $p(a_n) = 2^{-D}$ regardless of $n$. Hence, $S$ is normalized with respect to $S_{\text{max}}$, and the normalized string entropy $H$ is defined as $H = S/D$, where $0 \leq H \leq 1$ and $H = 1$ is obtained for completely random processes.

Let us consider the relationship between the string entropy and the Lyapunov exponent for a simple case,
i.e., a one-dimensional chaotic map $x_{n+1} = f(x_n)$. In this case, the Lyapunov exponent $\lambda$ is defined as

$$\lambda = \log | f'(x_0) | ,$$

where $f'$ denotes the derivative of $f$ with respect to $x$ and $x_0$ denotes the initial point. The Lyapunov exponent estimated using Eq. (4) is usually averaged over many initial points in the chaotic attractor (denoted as $\Omega$) to obtain the global Lyapunov exponent as

$$\lambda = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \log | f'(x_n) | ,$$

$$= \int_{\Omega} \mu(x) \log | f'(x) | \, dx ,$$

where $\mu(x)$ denotes the probability density function of $x$. The string entropy is estimated using the probability density function $p(a)$ of the alphabets coded in $D$ bits. Hence, the global Lyapunov exponent is related via Eq. (5) with the string entropy, since $p$ is the $D$ successive products of $\mu$ of the coarse-grained $x$ with the binary expression along a trajectory in $\Omega$.

A similar relationship holds for multi-dimensional chaotic maps and chaotic flows by considering the Jacobian of the chaotic dynamics.

3. Numerical Analysis and Discussion

We conducted numerical experiments on estimating $H$ with the parameter settings of $D = 7$, $N = 120000$, $Q = 20000$, $T = 1$ for the logistic map, tent map, the Lorenz equations, and the AL equations, where all numerical calculations were performed in double precision on a 32-bit machine. No particular methods were used to reduce the accumulation of roundoff errors.

The logistic map is defined as $x_{n+1} = ax_n(1-x_n)$, where $0 \leq x_n \leq 1$. Figures 1(a), (b), (c), and (d) show a typical example of the estimated histograms under randomly chosen initial points $(x_0)$ for $a = 3.95$, 3.98, 3.99, and 4, respectively. The threshold value was set to $x_c = 0.5$. The initial 5000 data points were discarded to eliminate the initial transient parts from the analysis. Estimates of $H$ are summarized in Table 1. At $a = 3.95$, 3.98, and 3.99, there are missing alphabets regardless of the choice of $x_0$. These missing alphabets can be the clue to identify the value of $a$. At $a = 4$, all alphabets appear with approximately equal probability and $H$ approaches unity.

The results for the tent map are shown in Fig. 2 and Table 1. The tent map is defined as $x_{n+1} = 1 - 2 | x_n - 0.5 |$. The convergence of $x_n$ to the fixed points of $x = 0$ and $x = 1$ was circumvented by restricting the domain of $x_n$ to $\epsilon \leq x_n \leq 1 - \epsilon$ with $\epsilon = 10^{-6}$. The threshold value was set to $x_c = 0.5$, and the initial 5000 data points were eliminated from the analysis. In contrast to the logistic maps, there are many missing alphabets, whereas the alphabet ‘0000000’ (‘0’ in the decimal expression) appears very frequently. These observations were independent of the choice of $x_0$. The estimated $H$ is considerably smaller than those for the logistic maps. Hence, the tent map cannot be used as a secure pseudorandom number generator.

The results for the Lorenz equations [17] are shown in Figs. 3(a) and (b), and in Table 1. The Lorenz equations are defined as a three-dimensional system of ordinary differential equations:

$$\dot{x} = \sigma (y - x) ,$$

$$\dot{y} = rx - y - xz ,$$

$$\dot{z} = -\beta z + xy ,$$

where $\sigma = 10$, $r = 28$, and $\beta = 8/3$. The equations were numerically integrated using the 4th-order Runge-Kutta method with a time width of $\Delta t = 0.01$. The initial conditions for $x$, $y$, and $z$ were given as pseudorandom numbers subject to the standard normal distribution. The initial 5000 numerical solutions were eliminated from the analysis. The threshold values were set to $x_c = 0$ and $y_c = 0$. For both $x$ and $y$, there are no missing alphabets and several alphabets appear more frequently than other alphabets. Estimates of $H$ for $x$ and $y$ are close to unity but slightly smaller than that for the logistic map with $a = 4$.

Our final case study is concerned with the AL equations defined as

$$\dot{x} = \sigma \left( \text{tr} \left( [M^{-1}]^2 y \right) - x \right) ,$$

$$\dot{y} = R x - M z x - y ,$$

$$\dot{z} = M y x - z ,$$

$$R = \begin{bmatrix} R_0 \end{bmatrix} M^2 \Phi \mathbf{W} ,$$

where $x$ is a dimensionless scalar variable, $y$ = diag($y_1$, ..., $y_N$, ..., $y_K$) and $z$ = diag($z_1$, ..., $z_n$, ..., $z_K$) are dimensionless $K \times K$ diagonal matrices, $\text{tr}()$ represents the diagonal sum of a matrix, $\sigma$ and $R_0$ are dimensionless parameters corresponding to the Prandtl and reduced Rayleigh numbers, respectively, and $M$ denotes the diagonal matrix given by $M = \text{diag}(m_1, \ldots, m_n, \ldots, m_K)$ with

<table>
<thead>
<tr>
<th>Dynamics</th>
<th>$H$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic map ($a = 3.95$)</td>
<td>0.8877</td>
</tr>
<tr>
<td>Logistic map ($a = 3.98$)</td>
<td>0.9486</td>
</tr>
<tr>
<td>Logistic map ($a = 3.99$)</td>
<td>0.9553</td>
</tr>
<tr>
<td>Logistic map ($a = 4$)</td>
<td>0.9994</td>
</tr>
<tr>
<td>Tent map</td>
<td>0.6625</td>
</tr>
<tr>
<td>Lorenz equations: $x$</td>
<td>0.9954</td>
</tr>
<tr>
<td>Lorenz equations: $y$</td>
<td>0.9961</td>
</tr>
<tr>
<td>AL equations: $x$</td>
<td>0.9991</td>
</tr>
<tr>
<td>AL equations: $y_{100}$</td>
<td>0.9984</td>
</tr>
</tbody>
</table>
and (d) $\alpha = 1$ (one time step).

![Figure 1](image1.png)

**Figure 1:** Histograms of alphabets for the logistic maps with (a) $\alpha = 3.95$, (b) $\alpha = 3.98$, (c) $\alpha = 3.99$, and (d) $\alpha = 4$. $D = 7$ and $T = 1$ (one time step).

$m_1 = 1$ and $m_2$ through $m_K$ randomly taking values of $m_n = n$ or $m_n = n + 0.5$. $M$ can be used as a secret key for cryptography [14]. For the definitions of the diagonal coefficient matrices $\Phi$ and $W$, see [14]. The bifurcation parameters $\sigma$, $R_0$, and $\phi$ were set to $\sigma = 25$, $R_0 = 3185$, and $\phi = 0.36$ [rad].

We numerically integrated the AL equations in a similar way to the Lorenz equations, except with a time width of $\Delta t = 4 \times 10^{-4}$ and $K = 101$ under a random setting of $M$ and the initial conditions of $x$, $y$, and $z$ given as pseudorandom numbers subject to the standard normal distribution. Time series $x$ and $y_{100}$ were obtained by discrete sampling of the numerical solutions with a sampling time of $T = 1$ (2500 time steps) and the threshold values $x_c = 0$ and $y_c = 0$.

![Figure 2](image2.png)

**Figure 2:** Histograms of alphabets for the tent map. $D = 7$ and $T = 1$ (one time step).

Figures 4(a) and (b) show typical examples of the histograms for $x$ and $y_{100}$, respectively. Estimates of $H$ are summarized in Table 1. For both $x$ and $y_{100}$, there are no missing alphabets. Estimates of $H$ for $x$ and $y_{100}$ are close to unity.

![Figure 3](image3.png)

**Figure 3:** Histograms of alphabets for (a) $x$ and (b) $y$ of the Lorenz model. $D = 7$ and $T = 1$ (100 time steps).

We next estimated $H$ as a function of the bifurcation parameter $R_0$ for $y_{100}$ of the AL equations. Figure 5 shows the estimates of $H$ as a function of $R_0$, where the sampling time interval $T$ was set to 0.01 (25 time steps) to reduce the total computational time for the numerical integration of the AL equations and $R_0$ was increased from 1000 to 3190 with an increment width of $\Delta R_0 = 10$. $H$ increases as $R_0$ increases, and approaches unity at $R_0 > 3000$, where the AL equations appear to generate fully developed chaotic time series.

In conclusion, our numerical analysis indicates that the string entropy method is capable of characterizing chaotic dynamics. In particular, the missing alphabets in the estimated histograms provide important features of the chaotic trajectories governed by the
when applying the chaotic dynamics to cryptography, $H$ should be as close as possible to its maximum value of unity, and there should be no missing alphabets. In this sense, the AL equations is useful for a pseudorandom generator for a one-time pad cipher. Recently, we have verified that the pseudorandom numbers generated by the AL equations pass the statistical tests of NIST 800-22, which will be reported in a future paper.

Acknowledgment

This study was partly supported by JSPS KAKENHI Grant Number JP15K00353.

References


Analysis on Differences of Japanese and English Languages by the Complex Network Theory

Mayumi Tatara†, Yutaka Shimada†‡, Kantaro Fujiwara†‡ and Tohru Ikeguchi†‡

†Department of Management Science, Graduate School of Engineering, Tokyo University of Science
‡Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6–3–1 Niijuku, Katsushika-ku, Tokyo, Japan

Abstract—The complex network based approaches considerably enhance our understanding of many real systems, for example, the Internet, human relations and neural networks. Languages can also be analyzed by the complex network based approach, because languages are described as a network consisting of words and their adjacency relations. Even though there are several researches on the language networks, they mainly focus on a specific language, and there are few researches comparing different languages from the viewpoint of complex networks.

In this paper, we generate the language networks from literature written in Japanese and English, and investigate differences of their network structures between Japanese and English. As a result, the structural properties of Japanese language networks are clearly different from those of English ones.

1. Introduction

Many natural, social, and artificial systems are described as networks which consist of a set of links and a set of nodes. The complex network theory has revealed common structural properties underlying the networks obtained from various types of real systems [1, 2]. Languages have also been analyzed from the viewpoint of complex networks. For example, Ref. [3] shows that the language networks describing co-occurrence of words have small-world and scale-free properties. The language networks have also been used as one of benchmarks for evaluating community detection methods [4]. In these previous studies, the language networks are generated from one specific language. In this paper, we raise a question whether we can quantify differences between one language and other languages from the viewpoint of network structures. To accomplish this issue, we generate the language networks from Japanese and English literature, and investigate differences between Japanese and English languages by analyzing their network structures.

2. Data

In this paper, we used Japanese literature provided from the web site “Aozora-bunko” [5] and English literature provided from the web site “Project Gutenberg” [6]. We choose 36 literature (18 each) which have higher access rankings in these websites [5, 6]. Tables 1 and 2 show authors and titles of the Japanese literature and the English literature that we used in this paper. We generated 36 language networks from these literature.

Table 1: Authors and titles of Japanese literature

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kenji Miyazawa</td>
<td>Ginga tetsudo no yoru</td>
</tr>
<tr>
<td>Ryunosuke Akutagawa</td>
<td>Imogayu</td>
</tr>
<tr>
<td>Soseki Natsume</td>
<td>Kokoro</td>
</tr>
<tr>
<td>Osamu Dazai</td>
<td>Hashire Merosu</td>
</tr>
<tr>
<td>Motojirō Kajii</td>
<td>Lemon</td>
</tr>
<tr>
<td>Nankichi Niimi</td>
<td>Gongitsune</td>
</tr>
<tr>
<td>Franz Kafka (Translated by Yoshito Harada)</td>
<td>Henshin</td>
</tr>
<tr>
<td>Katai Tayama</td>
<td>Futon</td>
</tr>
<tr>
<td>Mimei Ogawa</td>
<td>Akai rosoku to ningyo</td>
</tr>
<tr>
<td>Torahiko Terada</td>
<td>Kagakusha to atama</td>
</tr>
<tr>
<td>Ohgai Mori</td>
<td>Takasebune</td>
</tr>
<tr>
<td>Kyoka Izumi</td>
<td>Koyahijiri</td>
</tr>
<tr>
<td>Kotaro Takamura</td>
<td>Chieko no hansei</td>
</tr>
<tr>
<td>Juza Unno</td>
<td>Dauchu enseitai</td>
</tr>
<tr>
<td>Ango Sakaguchi</td>
<td>Mo gunbi ha iranai</td>
</tr>
<tr>
<td>Kunihiko Sugawa</td>
<td>Mujinto ni ikiru jurokunin</td>
</tr>
<tr>
<td>Sakutarō Hagiwara</td>
<td>Nekomachi</td>
</tr>
<tr>
<td>Kunio Yanagida</td>
<td>Yama no jinsei</td>
</tr>
</tbody>
</table>

3. Methods

3.1. How to generate language networks

In our study, we generated language networks by the following two methods.

Method 1 We defined nodes as words and links as the adjacency relation between the words, where each word connects with its nearest neighbors in the same sentence by the links. Figure 1(a) shows how to generate the language network by the method 1.
Table 2: Authors and titles of English literature

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lewis Carroll</td>
<td>Alice's Adventures in Wonderland</td>
</tr>
<tr>
<td>Mark Twain</td>
<td>The Adventures of Tom Sawyer</td>
</tr>
<tr>
<td>The Brothers Grimm</td>
<td>Grimm's Fairy Tales</td>
</tr>
<tr>
<td>J. M. Barrie</td>
<td>Peter Pan</td>
</tr>
<tr>
<td>Charles Dickens</td>
<td>A Christmas Carol in Prose; Being a Ghost Story of Christmas</td>
</tr>
<tr>
<td></td>
<td>A Tale of Two Cities</td>
</tr>
<tr>
<td>Jane Austen</td>
<td>Pride and Prejudice</td>
</tr>
<tr>
<td></td>
<td>Emma</td>
</tr>
<tr>
<td>Arthur Conan Doyle</td>
<td>The Adventures of Sherlock Holmes</td>
</tr>
<tr>
<td>Henrik Ibsen</td>
<td>A Doll's House</td>
</tr>
<tr>
<td>Jonathan Swift</td>
<td>A Modest Proposal</td>
</tr>
<tr>
<td>Daniel Defoe</td>
<td>The Life and Adventures of Robinson Crusoe</td>
</tr>
<tr>
<td>Mary Wollstonecraft</td>
<td>Frankenstein or The Modern Prometheus</td>
</tr>
<tr>
<td>Oscar Wilde</td>
<td>The Picture of Dorian Gray</td>
</tr>
<tr>
<td>Bram Stoker</td>
<td>Dracula</td>
</tr>
<tr>
<td>Lee Sutton</td>
<td>Venus Boy</td>
</tr>
<tr>
<td>Jack Sharkey</td>
<td>The Secret Martians</td>
</tr>
<tr>
<td>Robert Louis Stevenson</td>
<td>Treasure Island</td>
</tr>
</tbody>
</table>

Method 2  Each word connects with its next nearest neighbors in the same sentences. Figure 1(b) shows how to generate the language network by the method 2.

No links have weights and directions in this paper. Even if the same pairs of adjacent words occur more than once in the same literature, the number of links between these nodes is only one. In addition, symbols including punctuations and brackets are not contained in the language networks, and the words which are adjacent to these symbols are not connected with links. Self-loops such as “very very” are omitted from the networks.

To construct language networks for Japanese texts, we have to use a morphological analysis tool, because in Japanese texts, each word is not separated by a space. The morphological analysis enables us to automatically identify words from Japanese texts. In this paper, we used MeCab which is one of the morphological analysis tools for Japanese language [7].

I am Japanese.
I am in a graduate school.
I study complex networks.

I am Japanese.
I am in a graduate school.
I study complex networks.

(a) The method 1

(b) The method 2

Figure 1: How to generate language networks.
3.2. Measures of complex networks

We calculated the characteristic path length and the clustering coefficient of the language networks generated by the methods 1 and 2. The characteristic path length is the average of the shortest path lengths of all pairs of two nodes in the network. Let \( l_{ij} \) be the shortest path length from the node \( v_i \) to the node \( v_j \), and \( N \) be the number of nodes in the network. The characteristic path length \( L \) is then given by

\[
L = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} l_{ij}.
\]  

(1)

The clustering coefficient \( C \) is defined as follows. Let \( k_i \) be the degree of the node \( v_i \), and then at most \( k_i(k_i - 1)/2 \) links can exist between the adjacent nodes of \( v_i \). The ratio of the actual number of links between the adjacent nodes of \( v_i \) to the maximum number of such links is

\[
C_i = \frac{\text{the number of links between the adjacent nodes of } v_i}{k_i(k_i - 1)/2}.
\]  

(2)

The clustering coefficient \( C \) is then defined as

\[
C = \frac{1}{N} \sum_{i=1}^{N} C_i.
\]  

(3)

By calculating these values, we compared the network structures generated from the Japanese and English language networks. To compare the language networks with different sizes, \( L \) and \( C \) are normalized by \( L \) and \( C \) of randomized networks which are generated by randomizing links in the original language network so that the degree of each node are preserved [8]. In this randomization method, we first randomly selected two links which do not share nodes. Next, we selected one node from each of these links at random, and exchanged them. Repeating this procedure, we generated randomized networks.

4. Result

Table 3(a) shows the number of nodes and the number of links in the Japanese language networks, and Table 3(b) shows those in the English ones. In both Tables 3(a) and 3(b), \( M_1 \) indicates the number of links in the language network generated by the method 1, and \( M_2 \) indicates that by the method 2. From Tables 3(a) and 3(b), the number of links is about twice as large in the language networks generated from the method 1 as in those from the method 2 in both Japanese and English literature.

Figure 2 shows structural comparison between the Japanese and the English language networks by the normalized characteristic path length \( L_0/L_R \) and the normalized clustering coefficient \( C_0/C_R \), where \( L_0 \) is the characteristic path length of the original network, \( L_R \) is that of the randomized network, \( C_0 \) is the clustering coefficient of the original network, and \( C_R \) is that of the randomized network. From Fig. 2, the distribution of \( (C_0/C_R, L_0/L_R) \) is

Table 3: The numbers of nodes and edges in (a) Japanese language and (b) English language networks

(a)

<table>
<thead>
<tr>
<th>Title</th>
<th>( N )</th>
<th>( M_1 )</th>
<th>( M_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ginga tetsudo no yoru</td>
<td>2,586</td>
<td>9,386</td>
<td>18,457</td>
</tr>
<tr>
<td>Imogayu</td>
<td>1,854</td>
<td>5,135</td>
<td>9,781</td>
</tr>
<tr>
<td>Kokoro</td>
<td>6,617</td>
<td>30,196</td>
<td>61,588</td>
</tr>
<tr>
<td>Hashire Merosu</td>
<td>1,373</td>
<td>3,222</td>
<td>5,982</td>
</tr>
<tr>
<td>Lemon</td>
<td>866</td>
<td>1,954</td>
<td>3,673</td>
</tr>
<tr>
<td>Gongitsune</td>
<td>669</td>
<td>1,616</td>
<td>3,098</td>
</tr>
<tr>
<td>Henshin</td>
<td>3,220</td>
<td>12,531</td>
<td>25,184</td>
</tr>
<tr>
<td>Futon</td>
<td>3,844</td>
<td>13,281</td>
<td>25,841</td>
</tr>
<tr>
<td>Akai rosoku to ningyo</td>
<td>737</td>
<td>2,147</td>
<td>4,066</td>
</tr>
<tr>
<td>Kagakusha to atama</td>
<td>576</td>
<td>1,371</td>
<td>2,556</td>
</tr>
<tr>
<td>Takasebune</td>
<td>1,076</td>
<td>2,915</td>
<td>5,561</td>
</tr>
<tr>
<td>Koyahijiri</td>
<td>3,905</td>
<td>12,798</td>
<td>24,358</td>
</tr>
<tr>
<td>Chieko no hansei</td>
<td>1,816</td>
<td>4,761</td>
<td>9,103</td>
</tr>
<tr>
<td>Dauchu enseitai</td>
<td>3,392</td>
<td>12,657</td>
<td>24,964</td>
</tr>
<tr>
<td>Mo gunbi ha iranai</td>
<td>1,220</td>
<td>3,060</td>
<td>5,848</td>
</tr>
<tr>
<td>Mijinto ni iikiru jurokunin</td>
<td>5,018</td>
<td>21,337</td>
<td>42,663</td>
</tr>
<tr>
<td>Nekomachi</td>
<td>1,414</td>
<td>3,769</td>
<td>7,079</td>
</tr>
<tr>
<td>Yama no jinsei</td>
<td>8,308</td>
<td>31,767</td>
<td>63,166</td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Title</th>
<th>( N )</th>
<th>( M_1 )</th>
<th>( M_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alice’s Adventures in Wonderland</td>
<td>2,647</td>
<td>12,864</td>
<td>24,407</td>
</tr>
<tr>
<td>The Adventures of Tom Sawyer</td>
<td>7,499</td>
<td>35,086</td>
<td>65,533</td>
</tr>
<tr>
<td>Grimm’s Fairy Tales</td>
<td>4,939</td>
<td>34,920</td>
<td>66,559</td>
</tr>
<tr>
<td>Peter Pan</td>
<td>4,964</td>
<td>23,424</td>
<td>43,614</td>
</tr>
<tr>
<td>A Christmas Carol in Prose: Being a Ghost</td>
<td>4,365</td>
<td>16,405</td>
<td>30,600</td>
</tr>
<tr>
<td>Story of Christmas</td>
<td>10,150</td>
<td>58,772</td>
<td>108,891</td>
</tr>
<tr>
<td>A Tale of Two Cities</td>
<td>6,489</td>
<td>47,762</td>
<td>90,223</td>
</tr>
<tr>
<td>Emma</td>
<td>7,351</td>
<td>57,393</td>
<td>107,570</td>
</tr>
<tr>
<td>The Adventures of Sherlock Holmes</td>
<td>8,284</td>
<td>44,198</td>
<td>82,617</td>
</tr>
<tr>
<td>A Doll’s House</td>
<td>2,451</td>
<td>11,161</td>
<td>20,451</td>
</tr>
<tr>
<td>A Modest Proposal</td>
<td>1,075</td>
<td>2,716</td>
<td>5,220</td>
</tr>
<tr>
<td>The Life and Adventures of Robinson Crusoe</td>
<td>6,704</td>
<td>44,912</td>
<td>84,972</td>
</tr>
<tr>
<td>Frankenstein or The Modern Prometheus</td>
<td>7,092</td>
<td>37,587</td>
<td>70,462</td>
</tr>
<tr>
<td>The Picture of Dorian Gray</td>
<td>7,075</td>
<td>34,451</td>
<td>64,304</td>
</tr>
<tr>
<td>Dracula</td>
<td>9,701</td>
<td>60,779</td>
<td>113,058</td>
</tr>
<tr>
<td>Venus Boy</td>
<td>3,475</td>
<td>18,424</td>
<td>34,198</td>
</tr>
<tr>
<td>The Secret Martians</td>
<td>5,775</td>
<td>24,083</td>
<td>45,063</td>
</tr>
<tr>
<td>Treasure Island</td>
<td>6,166</td>
<td>32,359</td>
<td>60,579</td>
</tr>
</tbody>
</table>
classified into two classes corresponding to the Japanese language networks and English ones. In case of the language networks generated by the method 1 (Fig. 2(a)), the distribution of the Japanese language networks is located in the upper-left part, and that of the English language networks is located in the bottom-right part. However, in the case of the method 2 (Fig. 2(b)), the distribution of the Japanese language networks is located in the bottom-right part, and that of the English language networks is located in the upper-left part. According to Table 3, when we change the method for generating networks from the method 1 to 2, the number of links is equally doubled in almost all Japanese and English language networks. In spite of this fact, the changes of $L_O/L_R$ and $C_O/C_R$ in the Japanese language networks are larger than those in the English ones. These differences between Japanese and English language networks might be due to the difference of grammatical features between Japanese and English languages.

5. Conclusion

In this paper, we generated the networks from 36 literature written in Japanese and English, and investigated their network structures. As a result, the characteristic path lengths and the clustering coefficients of the Japanese language networks and the English language networks are classified into different classes. In addition, distribution tendency depends on how to generate networks. These differences might come from the grammatical features of Japanese and English languages.

Acknowledgments

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. is supported by Grant-in-Aid for Research Activity Start-up (No. 26880020) and by Grant-in-Aid for Young Scientists B (No. 16K16126) from JSPS. The research of K.F. was supported by Grant-in-Aid for Challenging Exploratory Research (No. 15K12137) and by Grant-in-Aid for Young Scientists B (No. 16K16138) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Challenging Exploratory Research (No. 24650116) from JSPS and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No. 15KT0112) from JSPS.

References

Irregularity of Inter-Event Interval of Diastrophism

Keisuke Miyata\textsuperscript{1}, Yutaka Shimada\textsuperscript{2}, Kantaro Fujiwara\textsuperscript{3} and Tohru Ikeguchi\textsuperscript{4,5}

\textsuperscript{1}Department of Management Science, Graduate School of Engineering, \textsuperscript{2}Department of Information and Computer Technology, Faculty of Engineering, \textsuperscript{3}Tokyo University of Science

Abstract—One of the main factors of earthquake is stress increase due to mutual moving processes of base rocks in a plate deep beneath the earth’s surface. Recently, the seismic activity of diastrophism has been analyzed. Using GPS-based control station data by converting into inter-event interval (IEI) of the diastrophism occurrence, we investigated its irregularity using several measures for analyzing inter-spike intervals of neural signals. As a result, we confirmed that the IEI irregularity statistics increase before large-scale earthquakes occur.

1. Introduction

Earthquakes occur during the process of the fluctuation in deep part of plates constituting the earth. Plural marine plates sink under plural continental plates around the Japanese Islands. Complicated forces are driven by plural plates. Then, Japan is one of the most eminent earthquake-prone zone in the world. Therefore, various methods are proposed to predict when, where and how large earthquake occurs\textsuperscript{[1, 2]}. It has been reported that very small size diastrophism is observed by GPS\textsuperscript{(Global Positioning System)-based control station just before large-scale earthquakes occur\textsuperscript{[3]}. The GPS-based control station is a device that receives the information of the position coordinate from the satellites using GPS. In Japan, there exist approximately 1,300 stations. They located approximately every 20km, and these stations are managed by Geospatial Information Authority of Japan.

2. Data

In this paper, we used a daily value of GPS-based control station data offer service offered by Geospatial Information Authority of Japan\textsuperscript{[4]}. Daily position coordinate information in these data consist of the information of the GPS-based control station in all over Japan. Among various data, we used a three-dimensional coordinate values as shown in Fig. 1. The direction from the center of gravity to the intersection Greenwich meridian and the equator is set as the X-axis, the direction from the center of gravity to the intersection of 90\textdegree meridian east the equator is set as the Y-axis, and the direction from the center of gravity to the North Pole is set as the Z-axis. Figure 2 shows three-dimensional coordinate values from January 1st, 1999 until April 16th, 2016, observed in Matsunoyama\textsuperscript{1}. From Fig. 2, we can find that these coordinate values fluctuate every day and they suddenly show a large fluctuation on March, 2011. This is because of the Great East Japan Earthquake Disaster occurred in March 11th, 2011.

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure1.png}
\caption{Schematic illustration of the three-dimensional coordinate\textsuperscript{[5]}.}
\end{figure}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{figure2.png}
\caption{Time series of three-dimensional daily coordinate values from January 1st, 1999 until April 16th, 2016, observed in Matsunoyama. The horizontal axis is time [yyyy/mm] and the vertical axis is the coordinate values of each axis.}
\end{figure}

\footnotesize\textsuperscript{1}37°07′91″ North latitude, 138°60′87″ East Longitude
3. Method

We transformed the three-dimensional coordinate values to a distance of the movement. We defined the distance \(d_t\) by Eq. (1):

\[
d_t = \sqrt{(x_t - x_{t-1})^2 + (y_t - y_{t-1})^2 + (z_t - z_{t-1})^2}
\]  

(1)

Namely, the distance \(d_t\) is defined as difference between coordinates \((x_{t-1}, y_{t-1}, z_{t-1})\) of the day \(t-1\) and \((x_t, y_t, z_t)\) of the day \(t\) with three-dimensional coordinates values. Figure 3 shows a time series of the distance \(d_t\) using the three-dimensional coordinate values as defined by Fig. 2. From Fig. 3, we can find that these coordinate values fluctuate every day, however the values suddenly show a great fluctuation on March, 2011, which is the same tendency as shown in Fig. 2.

![Figure 3: A time series of \(d_t\) from January 1st, 1999 until April 16th, 2016 observed in Matsunoyama. The horizontal axis is time [yyyy/mm] and the vertical axis is the distance \(d_t\) [m] defined by Eq. (1).](image)

To analyze the data, we transformed the distance time series \(d_t\) to a point process. Namely, by setting a threshold \(\theta\) for the distance \(d_t\) which we defined in Eq. (1), we generated a point process in which an event is defined as a time when \(d_t\) exceeds the threshold \(\theta\).

Using the obtained point process, we investigated an irregularity of the inter-event interval (IEI) using two statistics: \(LV[6]\) and \(IR[7]\). These statistics are used to analyze the inter-spike interval (ISI) in the field of neuroscience. The indices, \(LV\) and \(IR\), quantify local variations of IEI, defined as follows:

\[
LV = \frac{3}{n-1} \sum_{i=1}^{n-1} \frac{(X_i - X_{i+1})^2}{(X_i + X_{i+1})^2},
\]

(2)

\[
IR = \frac{1}{n-1} \sum_{i=1}^{n-1} \log X_i - \log X_{i+1},
\]

(3)

where \(X_i(i = 1, 2, \cdots n)\) is the \(i\) th IEI and \(n\) is the total number of IEI.

It is possible to quantify the deviation using a statistic such as the coefficient of variation (CV). However, CV depends on the occurrence rate of IEI and it is not suitable for non-stationary data analysis. Thus, we analyzed \(LV\) and \(IR\) that do not depend on the occurrence rate of IEI.

If the observed point process obeys a Poisson process, \(LV\) takes 1 and \(IR\) takes 2 log 2. On the other hand, both \(LV\) and \(IR\) take 0 for perfectly a regular point process. To obtain reliable results, it is inevitable to have a larger number of IEI. Then, in this paper, we set the number of IEI as \(n = 100\).

4. Results

We investigated the earthquake occurred in Niigata Prefecture, Japan, October 23th, 2004. This earthquake is called the Niigata Prefecture Chuetsu earthquake. This earthquake was a large-scale earthquake whose magnitude is 6.8 and shindo is 7, which is a Japanese intensity scale. We used the data observed in Matsunoyama and Izumozaki \(^2\) which is located within the range of 50km from the epicenter (Fig. 4).

![Figure 4: The observation points which we used for analyzing the data of the Niigata Prefecture Chuetsu earthquake. The epicenter is marked by star and ■ is the observation point: (a) Matsunoyama and (b) Izumozaki. The blue circle represents the region of radius 50km from the epicenter. This map is made based on [4].](image)

![Figure 5: Temporal changes of (a) \(LV\) and (b) \(IR\) from January 1st, 1999 until April 16th, 2016 at the observation point of Matsunoyama (\(\theta = 0.01, n = 100\)). The horizontal axis is time [yyyy/mm] and the vertical axes are (a) \(LV\) and (b) \(IR\).](image)

\(^2\)37°53'58" North latitude, 138°70'69" East Longitude
Figure 5 shows the temporal change of \( LV \) and \( IR \) obtained from IEI with \( \theta = 0.01, n = 100 \) in Matsunoyama. Figure 6 shows the temporal change of \( LV \) and \( IR \) obtained from IEI with \( \theta = 0.01, n = 100 \) in Izumozaiki. Blue vertical bars in Figs. 5 and 6 show dates of the occurrence of the main shock. As shown in Fig. 5, it is also revealed that \( LV \) and \( IR \) suddenly rise from March, 2003 to March, 2004 before the occurrence of the main shock. From Fig. 6, it is also revealed that \( LV \) and \( IR \) suddenly rise from January, 2002 through March, 2004 before the occurrence of main shock. These results indicate that irregularity of the IEI of the diastrophism increases while a short time in these observation points.

Next, we investigated the behavior of each statistic when we changed the threshold of these data. The data consist of daily coordinate values. Therefore, if we set the threshold \( \theta = 0 \), these statistics \( LV \) and \( IR \) become 0 because point process becomes periodic. Therefore, it becomes hard to observe the change of the IEI statistics when we set the threshold too low. On the other hand, it becomes hard to get a sufficient number of IEIs when we set the threshold too high. In Figs. 5 and 6, the mean of \( d_i \) of Matsunoyama is 0.009354 and the mean of \( d_i \) of Izumozaiki is 0.008536. Then, we set the threshold \( \theta = 0.01 \) which is close to the mean of \( d_i \). In the following, we will investigate the behaviors of \( LV \) and \( IR \) by changing the threshold \( \theta \) from 0.01.

Figures 7 and 8 show the temporal changes of \( LV \) and \( IR \) when we changed the threshold \( \theta \) from 0.005 to 0.015 in Matsunoyama and Izumozaiki. Thick red lines show the statistics in case that \( \theta = 0.01 \). As shown in Figs. 7 and 8, statistics increase as the value of \( \theta \) increases. When the threshold \( \theta \) increases, only large diastrophism is detected and the obtained point process would be sparse, which makes difficult to visualize the change of the statistics with high time resolution. In addition, in the case that \( \theta = 0.005 \), the statistics fluctuate with almost constant small values and no rapid increase. Because the point process becomes regular by decreasing the threshold \( \theta \), \( LV \) and \( IR \) become low.
Figure 9: Temporal change of (a)LV and (b)IR from January 1st, 1999 until April 16th, 2016 at the observation point of Matsunoyama when $\theta$ is changed from 0.015 to 0.025 by 0.0025 ($n = 100$). The horizontal axis is time [yyyy/mm] and the vertical axes are (a) LV and (b) IR.

Figure 10: Temporal change of (a)LV and (b)IR from January 1st, 1999 until April 16th, 2016 at the observation point of Izumozaki when $\theta$ is changed from 0.015 to 0.025 by 0.0025 ($n = 100$). The horizontal axis is time [yyyy/mm] and the vertical axes are (a) LV and (b) IR.

Figures 9 and 10 show LV and IR when we changed $\theta$ from 0.015 to 0.025 in Matsunoyama and Izumozaki. As shown in Figs. 9 and 10, LV and IR take high values generally and they become smaller when the threshold is higher. This is because the number of event decreases by increasing $\theta$. According to the above results, it is indicated that appropriate value of the threshold which is close to the mean of $d_i$ is necessary for visualizing characteristic behaviors of the statistics.

5. Conclusion

In this paper, we analyzed the interval of the change of three-dimensional coordinate values provided by GPS-based control station, using statistics of LV and IR that are often used for analyzing a point process in the field of neuroscience.

In Ref.[8], LV was used to analyze irregularity of IEIs transformed from time, spatial coordinate and magnitude of earthquake. Different from Ref.[8], we investigated temporal changes of the statistics of IEIs of the diastrophism observed from GPS-based control station data[4].

We analyzed the data of Niigata Prefecture Chuetsu earthquake at observation points located within the range of 50km from the epicenter. As a result, we showed that temporal changes of the statistics LV and IR were detected before the main shock occurrence. Moreover, we changed the value of the threshold and investigated how these statistics change. It is indicated that if the threshold values are set appropriately, large-scale earthquakes can be detected from the point processes.

Acknowledgments

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. is supported by Grant- in-Aid for Research Activity Start-up (No. 26880020) from JSPS. The research of K.F. was supported by Grant-in-Aid for Challenging Exploratory Research (No 15K12137) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Challenging Exploratory Research (No. 24650116) from JSPS and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No.15KT0112) from JSPS.

References

A spiking neuron model with two slopes and triangular wave base signal

Yusuke Matsuoka†

†Department of Electrical and Computer Engineering, National Institute of Technology, Yonago College, Hikona-cho 4448, Yonago-shi, Tottori, 683-8502, Japan. Email: ymatuoka AT yonago-k.ac.jp

Abstract—This paper considers a spiking neuron model with two slopes and triangular wave base signal. The state variable of this neuron repeats a integrate-and-fire dynamics and the slope of the state is alternatively changed at every resetting. We focus on the timing of the spike phase and derive the spike phase map. Using the map, we consider behavior of the neuron. It is shown that the spiking neuron exhibits not only chaos and periodic phenomena but also the superstable and co-existence phenomena.

1. Introduction

This paper considers a spiking neuron model with two slopes and triangular wave base signal. The state variable of this spiking neuron operates a integrate-and-fire dynamics. The state has a constant slope and rises with time. If the state reaches the threshold line, the neuron outputs a spike and the state resets to the triangular wave base signal where the resetting and output are instantaneous and simultaneous. Next, the state rises with another constant slope. After reaching the threshold and resetting to the base signal, the state again rises with the first slope. Repeating above dynamics, the neuron exhibits nonperiodic and periodic trajectory and corresponding to the output of the spike.

In order to consider the generating phenomenon, we focus on the timing of the spike and derive the spike phase map. Since this map is the return map and piecewise linear, it has advantage that we can analyze rigorously. When the parameters vary, the shape of the map, the slope of the segment and the number of discontinuous points change. Using the map and Lyapunov exponent, we show that the neuron generates various phenomena. In addition, the map exhibits the superstable periodic orbit (SSPO) [1] with various period and co-existence phenomena of chaos and SSPO. These phenomena are also shown, using the map.

Many papers have been studied artificial neuron models [2]-[4] and the spiking neuron is a kind of the model. The spiking neuron exhibits various nonlinear phenomena as shown in this paper though a simple model. Ref. [5] has studied for bifurcation phenomena. Using the spike-output and two neurons, the pulse-coupled system has been studied in Ref. [6]. These investigation results are basic for consideration of synchronization phenomenon. In addition, applications for the analog-to-digital converters have been considered to encode the timing of the spike [7]-[8]. We think that discussion and consideration in this paper are approach to advance in respect of above investigation.

\[
\begin{align*}
  b(\tau + 1) &= b(\tau), \\
  b(\tau) &= \begin{cases} 
    -k(\tau - \frac{1}{4}) - 1 & \text{for } 0 \leq \tau < 0.5, \\ 
    k(\tau - \frac{1}{4}) - 1 & \text{for } 0.5 \leq \tau < 1.
  \end{cases}
\end{align*}
\]

Previous work [5] has studied the spiking neuron model with triangular wave base signal. However the slope of the state in Ref. [5] is fixed one constant slope.

2. Basic dynamics and typical phenomena of the neuron

The dynamics of targeted spiking neuron in this paper is described by Equation (1).

\[
\begin{align*}
  \frac{dx}{dt} &= s_1, \quad y = 0 \quad \text{for } x < 0 \text{ and } l \text{ is even}, \\
  \frac{dx}{dt} &= s_2, \quad y = 0 \quad \text{for } x < 0 \text{ and } l \text{ is odd}, \\
  x(\tau +) &= b(\tau+), \quad y(\tau+) = 1, \quad \text{if } x \geq 0,
\end{align*}
\]

where \( \tau, x \) and \( y \) are dimensionless time, state variable and output of the neuron. Let \( l \) be the number of times at which the state hits threshold level \( x = 0 \) and \( l \) is nonnegative integer. The base signal \( b(\tau) \) is triangular waveform with period 1 as follows.

Figure 1: Basic dynamics of the spiking neuron.

Figure 2: An example of the waveform and output. \( s_1 = 2.4, s_2 = 1.7, k = 3.7 \).
This neuron has three parameters: slopes of the state variable $s_1, s_2$ and slope of the base signal $k$. $s_1, s_2, k$ are real parameters and we restrict to $0 < s_1, 0 < s_2, 0 < k < 4$, and the initial state $-1 < x(0) < 0$. An example of the dynamics is shown in Fig. 1. At first, the state $x(t)$ starting from the initial state $x(0) = x_0$ rises with the slope $s_1$. If $x$ reaches the threshold line, $x = 0$, the spike $y$ is outputted and the state $x$ resets to the base signal $b(t)$. For simplicity, we assume that the spike-output and resetting are instantaneous and simultaneous. Next, $x$ rises with another slope $s_2$. If $x$ again reaches the threshold line, the spike is outputted and $x$ resets to the base $b(t)$ instantaneously. The slope of the state is alternately changed to $s_1$ and $s_2$ at every resetting. The spiking neuron repeats above dynamics. Fig. 2 shows an example of the time waveform. We note that Ref. [5] have discussed the spiking neuron with fixed one constant slope $(s = s_1 = s_2)$ and studied for bifurcation phenomena.

In order to consider the behavior of this neuron, we define the spike position map. As shown in Fig. 1, let $\tau_n$ and $\tau_{n+1}$ (or $\tau'_n$ and $\tau'_{n+1}$) be $n$-th and $n+1$-st spike positions of odd numbers (or even numbers). The spike position $\tau_n$ determines next spike position $\tau'_n$ and $\tau'_n$ determines the spike position $\tau_{n+1}$. Since $\tau_{n+1}$ depends on $\tau_n$, the spike position map $F$ can be defined as follows.

$$
\begin{align*}
\tau_{n+1} &\equiv F(\tau_n) = F_1(F_2(\tau_n)), \\
\tau_{n+1} &\equiv F_1(\tau'_n) = \tau_n - \frac{1}{s_1}b(\tau'_n), \\
\tau'_n &\equiv F_2(\tau_n) = \tau_n - \frac{1}{s_2}b(\tau_n),
\end{align*}
$$

(3)

where functions $F_1$ and $F_2$ are described theoretically and are the same as those of the system in Ref. [5]. An example of the spike position map is shown in Fig. 3. This map is piecewise linear and the shape is varied by changing the parameters.

Here we introduce the phase $\theta$ as new variable and define the spike phase $\theta_n \equiv \tau_n \mod 1$ and $\theta'_n \equiv \tau'_n \mod 1$. That is, $\theta_n$ and $\theta'_n$ denote decimals of $\tau_n$ and $\tau'_n$. We can consider the same as the spike position map for the spike phase because $\theta_{n+1}$ depends on $\theta_n$. Therefore, we define the spike phase map $f$ as follows.

$$
\begin{align*}
\theta_{n+1} &= f(\theta_n) = g_1(\theta'_n) \mod 1, \\
\theta'_n &= g_2(\theta_n) \mod 1.
\end{align*}
$$

(4)

Functions $g_1$ and $g_2$ are described theoretically and are the same as those of the system in Ref. [5]. Fig. 4 shows examples of the spike phase map. Although the spike phase map is piecewise linear, discontinuous points appear because of taking modulus for the spike phase. The parameters of Fig. 4 (a) correspond to those of Fig. 3.

In order to consider behavior of the spike phase map, we describe some definitions as follows.

**Definition1:** A point $\theta_n$ is said to be a fixed point if $\theta_n = f(\theta_n)$. A point $\theta_n$ is said to be a periodic point with period $k$ if $\theta_n = f^k(\theta_n), \theta_n \neq f^j(\theta_n)$ for $1 \leq j < k$ where $f^k$ denotes the $k$-fold composition of $f$ and $k > 1$. The periodic points with period $k$ are said to be unstable (or stable) for initial state if $|Df^k(\theta_n)| > 1$ (or $|Df^k(\theta_n)| < 1$), where $Df \equiv \frac{df}{d\theta}$. An orbit of the stable periodic points is referred to as stable periodic orbit. If $f(I) \subseteq I$ and there exists some positive integer $l$ such that $|Df^l(\theta_n)| \geq 1$ for almost all $\theta_n \in I$ where $f \equiv [0, 1]$, then the orbit is unstable and the map $f$ is said to generate chaos [9].
Fig. 4 (a) and (b) exhibit chaotic orbits. Fig. 4 (c) has a fixed point (black circle) and exhibits periodic orbit with period 1. Fig. 4 (d) exhibits periodic orbit with period 7. The map exhibits a variety of behavior by changing the parameters.

We show bifurcation diagram and Lyapunov exponent $\lambda_n$ for $s_2$ in Fig. 5 when $s_1 = 2.4$ and $k = 3.7$ are fixed, where this figure corresponds to Fig. 4 (a) and (b). In these parameters, The range which has positive Lyapunov exponent is wide: the range generating chaotic orbit is wide. As shown in Fig. 4 (a) and (b), when the parameter $k$ is large, the slope of the segment on the map becomes steep and the orbit tends to be unstable. Therefore, the map tends to exhibit chaos.

Possible range which value of $\theta_n$ takes is wide in the neighborhood of $s_2 = 1.5$, however possible range is narrow in the neighborhood of $s_2 = 3.5$. $\theta_n$ is timing of the spike-output. In the neighborhood of $s_2 = 1.5$, it is shown that spike timings spread out in the range $[0, 1]$. In the neighborhood of $s_2 = 3.5$, it is shown that spike timings concentrate in a local region. Although both these phenomena are qualitatively chaotic, we can see that characteristics of the spike-output are different. In the spiking neuron model, the analog input is encoded through the spike timing and applications to the analog-to-digital converter have been studied in Refs. [7] and [8]. We think that investigation of the spike phase and spike-output characteristics are important but it is in future works.

Fig. 6 shows bifurcation diagram and Lyapunov exponent $\lambda_n$ for $s_2$ when $s_1 = 2.4$ and $k = 1.7$ are fixed, where this figure corresponds to Fig. 4 (c) and (d). When $k = 1.7$, the map has some parts of gradual slope as shown in Fig. 4 (c) and (d) because $k$ is small. Therefore, the orbit tends to become stable and tends to exhibit periodic orbit. In addition, various periodic orbits exist by changing the parameters and the range which has negative Lyapunov exponent is wider as compared with Fig. 5.

3. Various interesting phenomena

In Section 2, we show that this spiking neuron exhibits a variety of chaotic and periodic phenomena. In this section, more interesting phenomena are shown in using the spike phase map.

3.1. Superstable periodic orbits (SSPOs)

If the parameters $s_1 = k$ or $s_2 = k$ is fixed, the spike phase map has a flat part: there is the segment with the slope 0. Typical spike phase map is shown in Fig. 7. Here we describe the definition in respect of above as follows.

**Definition2:** The periodic points with period $k$ are said to be superstable for initial state if $|Df^n(\theta_n)| = 0$. An orbit of the superstable periodic points is referred to as superstable periodic orbit (SSPO).

In Fig. 7 (a), the map has a fixed point with the slope 0 and therefore exhibits the superstable fixed point (or SSPO...
with period 1). We note that Lyapunov exponent is negative infinity when the map exhibits the SSPO. In the steady state for Fig. 7 (b), we can see that the orbit starting from the flat segment returns to the flat segment. This map exhibits the SSPO with period 3. This neuron can generate various SSPOs by changing the parameters.

### 3.2. Various co-existence phenomena

Although the parameters of the system are same, for different initial state, the system exhibits different behavior in the steady state. This phenomenon is called co-existence phenomenon. Examples of the co-existence phenomenon are shown in Fig. 8. Fig. 8 (a) and (b) are the same parameters \( s_1 = 2.55, s_2 = 2.7 \) and \( k = 3.7 \), however the initial states are different. The initial state is \( x(0) = -0.8 \) in Fig. 8 (a) and is \( x(0) = -0.3 \) in Fig. 8 (b). We can see that Fig. 8 (a) exhibits chaotic orbit and Fig. 8 (b) exhibits periodic orbit (stable fixed point) in the steady state. Therefore, this neuron exhibits co-existence phenomenon of chaotic and periodic orbit for \( s_1 = 2.55, s_2 = 2.7 \) and \( k = 3.7 \). Fig. 8 (c) and (d) also have the same parameters \( s_1 = 2.4, s_2 = 3.7 \) and \( k = 3.7 \) and the initial states are different. We can see that Fig. 8 (c) exhibits chaotic orbit, however the orbit in Fig. 8 (b) hits the flat segment and the SSPO with period 2 is generated. In this case, this neuron exhibits co-existence phenomenon of chaos and SSPO for \( s_1 = 2.4, s_2 = 3.7 \) and \( k = 3.7 \). In Fig. 7 (a), we can see that the map has the superstable fixed point (black circle) while also has stable fixed point in the lower left part. In this case, this neuron exhibits co-existence phenomenon of the SSPO and periodic orbit. Proposed our spiking neuron can exhibit various co-existence phenomena and we think that very complicated bifurcation phenomena occur.

### 4. Conclusions

In this paper, a spiking neuron model with two slopes and triangular wave base signal has been studied. We derive the spike phase map and consider generating phenomena. By changing the parameters the spiking neuron exhibits chaos and periodic phenomena. We can see that the neuron exhibits various SSPOs and co-existence phenomena for some parameters. In the future, we intend to analyze bifurcation phenomena and consider characteristics of the spike-output.

### References


A Switching Ripple Reduction Technique for Current-Controlled 1-Dimensional DC/DC Boost Converter

Hiroyuki Fukue† Takuji Kousaka‡ Tetsushi Uta† Daisuke Ito** Hirokazu Ohtagaki† Hiroyuki Asahara‡

†Faculty of Engineering, Okayama University of Science, 1-1 Ridai-cho, Kita-ku, Okayama-shi, Okayama 700-0005, Japan.
‡Faculty of Engineering, Oita University, 700 Dannoharu, Oita-shi, Oita 870-1192, Japan.
* Center for Administration of Information Technology, Tokushima University, 2-1, Minami-Josanjima-cho, Tokushima-shi, Tokushima 770-8506, Japan.
** Faculty of Engineering, The University of Shiga Prefecture, 2500, Hassaka-cho, Hikone-shi, Shiga 522-8533, Japan.
Email:fukue@nonlineargroup.net, takuji@oita-u.ac.jp, ueta@tokushima-u.ac.jp, ito.d@e.usp.ac.jp, ohtagaki@ee.ous.ac.jp, asahara@ee.ous.ac.jp

Abstract—In this paper, we demonstrate a switching ripple reduction technique for current-controlled one-dimensional DC/DC boost converter. First, we show the circuit model, and then we explain behavior of the waveform. Next, we define the discrete map by sampling the inductor current at clock interval. Based on the discrete map, we derive a controlling gain. Finally, we try to reduce switching ripple of inductor current, and confirm the effectiveness of our controlling method.

1. Introduction

Power conversion circuits including switching devices have an interrupted characteristics due to the switching events, and rich nonlinear phenomena can be observed depending on the circuit parameter [1, 2]. It is important to analyze the bifurcation parameter for designing suitable circuit parameters.

In this study, we focus on the current-controlled DC/DC boost converter. The switching action depends on its own state and a time interval. This circuit exhibits rich nonlinear phenomena including chaotic behavior depending on the circuit parameter, and it is known that the switching ripple becomes minimum when it behaves with period-one oscillation. Therefore, we can say that we should design the suitable circuit parameters for making the circuit behave with period-one oscillation. On the other hand, even if the circuit exhibits chaotic behavior, we can control the chaotic behavior to the unstable period-one oscillation by using the chaos controlling technique. There are many chaos controlling techniques [3, 4, 5, 6], and one of which is our proposed technique [6].

This paper addresses to demonstrate application of our proposed chaos controlling technique to DC/DC converter. First, we show the circuit model, and then we explain behavior of the waveforms. Next, we define the discrete map, and derive a controlling gain. Finally, we try to reduce switching ripple of inductor current.

2. DC/DC boost converter and its control algorithm

Figure 1 shows the DC/DC boost converter, where the circuit parameters are

\[ R = 70[\Omega], \quad L = 1[mH], \quad C = 100[\mu F], \]

\[ r = 1[\Omega], \quad T = 17[\mu s], \quad I_{ref} = 0.92[\text{A}], \quad E = 15[\text{V}]. \] (1)

Note that the capacitor voltage is constant under this circuit parameter, and we show the capacitor voltage as \( E_0 \) in the following. Therefore, the circuit equation is described as follows.

\[ L \frac{di}{dt} = \begin{cases} E - ri, & \text{for switch is ON} \\ E - E_0 - ri, & \text{for switch is OFF} \end{cases} \] (2)

The solution of Eq. 2 is

\[ i(t) = \begin{cases} (i_0 - \frac{E}{r}) e^{-\frac{t}{r}} + \frac{E}{r}, & \text{for switch is ON} \\ (i_0 - \frac{E - E_0}{r}) e^{-\frac{t}{r}} + \frac{E - E_0}{r}, & \text{for switch is OFF} \end{cases}, \] (3)

where \( i_0 \) is an initial value at \( t = 0 \). Here, \( i_R \) is given by

\[ i_R = \frac{E_0}{r}. \] (4)

The transfer factor is

\[ M = \frac{E_0}{E} = \frac{1}{D'} \]. (5)

![Figure 1: DC/DC boost converter](image-url)
Therefore, we have
\[ D' = \frac{E}{E_0}, \]  
(6)
where \( D + D' = 1 \), and \( D \) is duty ratio. In addition, it holds the following equation:
\[ i = \frac{E_0}{RD'} = \frac{E_0}{R} \frac{1}{D'} = i_R \frac{E_0}{D'}, \]  
(7)
and we get
\[ i_R = iD'. \]  
(8)
Based on Eq. (4), Eq. (6), and Eq. (8), \( i_R \) is rewritten as follows:
\[ i_R = iD' = \frac{E_0}{R} = \frac{i}{E_0}. \]  
(9)
By solving Eq. (9), we get
\[ E_0 = \sqrt{iER}. \]  
(10)
In the following, we assume \( i = I_{ref} \), where \( I_{ref} \) is a reference value.

We now explain switching rule and its control algorithm. The current-control is applied to turn on, or turn off, the switch. We assume that the switch is ON at first position at \( t = 0 \), and inductor current increases. If the inductor current reaches the reference value \( I_{ref} \) at \( t = t_{on} \), the switch turns off and inductor current decreases. When the clock pulse is applied at \( t = T \), the switch turns on.

We control the switching action based on the algorithm proposed in Ref. [6]. Therefore, the switch turns on at \( t = nT + t_{on} + \Delta t_{on} \), where \( n = 1, 2, 3, \cdots \), and \( \Delta t_{on} \) is defined as follows:
\[ \Delta t_{on} = k\Delta t_n. \]  
(11)
In Eq. (11), \( k \) is a controlled gain described by
\[ k = -\frac{di_{n+1}}{dt_{on}}, \]  
(12)
and \( \Delta t_n \) is a perturbation described by
\[ \Delta t_n = i_n - i^*_n, \]  
(13)
where \( i_n \) is inductor current at \( t = nT \), \( i_{n+1} \) is inductor current at \( t = (n + 1)T \), and \( i^*_n \) is fixed point.

3. Demonstration of application

Using the algorithm proposed in Ref. [6], and controlling the inductor current to unstable period-one oscillation, we try to reduce switching ripple smaller.

The discrete map is needed for deriving control gain, and it is illustrated in Fig. 3. In case-1, the switch keeps ON during clock interval, and the discrete map \( i_{n+1} \) is described as follows:
\[ i_{n+1} = \left( i_n - \frac{E}{r} \right) e^{-\frac{t}{T}} + \frac{E}{r}. \]  
(14)
On the other hand, switch turns off at \( t = nT + t_{on} \) in case-2. Therefore, the discrete map \( i_{n+1} \) is described as follows:
\[ i_{n+1} = \left( i_n - \frac{E - E_0}{r} \right) e^{-\frac{t}{T}} + \frac{E}{r}. \]  
(15)

The differentials of Eq. (14) and Eq. (15) with respect to \( i_n \) are described as follows:
\[ \frac{di_{n+1}}{di_n} = \begin{cases} e^{-\frac{t}{T}}, & \text{for case-1} \\ \frac{r}{L} I_{ref} - \frac{E - E_0}{r} e^{-\frac{t}{T}}, & \text{for case-2} \end{cases}. \]
(16)
Likewise, the differentials of Eq. (14) and Eq. (15) with respect to \( t_{on} \) are described as follows:
\[ \frac{di_{n+1}}{dt_{on}} = \begin{cases} 0, & \text{for case-1} \\ \frac{r}{L} I_{ref} - \frac{E - E_0}{r} e^{-\frac{t}{T}}, & \text{for case-2} \end{cases}. \]
(17)
Because the fixed point satisfies \( i_n = i_{n+1} = i^*_n \), it is defined from Eq. (15) as follows:
\[ i^*_n = \frac{E}{r} \left( I_{ref} - \frac{E}{r} - E_0 \right) e^{-\frac{t}{T}}. \]
(18)
Therefore, the control gain \( k \) is given by
\[ k = \frac{r}{L} I_{ref} - \frac{E}{r} e^{-\frac{t}{T}}. \]
(19)

Figure 3 shows waveform of inductor current; (a) is overall view, where we start to control around \( t = 0.9\) ms, (b) is enlarged view of uncontrolled waveform, (c) is enlarged view of controlled waveform, and (d) is \( \Delta t_{on} \). We observe period-two oscillation under the circuit parameter shown in Eq. (1) (see Fig. 3(b)). On the other hand, if we apply the
control method discussed in Sec. 2, the period-two oscillation is controlled to unstable period-one oscillation (see Fig. 3(c)). It is clear from Fig. 3(d) that the control is completed around $t = 6\,[\text{ms}]$, because $\Delta t_{\text{on}} = 0$. We observe that the switching ripple is reduced by using the control method. We know that the circuit exhibits rich nonlinear phenomena by changing the circuit parameter. The switching ripple reduction technique proposed in Ref. [6] is applicable period-two or more oscillation including chaos, and can reduce the switching ripple.

4. Conclusion

We demonstrated a switching ripple reduction technique for current-controlled one-dimensional DC/DC boost converter. First, we showed DC/DC boost converter, and then we explain the switching events. Next, we introduced a control method of switching action reported in Ref. [6]. Using discrete map, which is a sampling inductor current at clock interval, we derive a control gain. Finally, we tried to reduce the switching ripple of inductor current, and confirmed the effectiveness of the method.

The control algorithm is effective only for one-dimensional DC/DC converters. In future, we improve the control algorithm for $n$-dimensional interrupted circuits, and confirm validity of the method.

References


Chaos near the Chenciner bubbles from a piecewise-constant system

Tri Quoc TRUONG\textsuperscript{a}, Tadashi TSUBONE\textsuperscript{a}, Naohiko INABA\textsuperscript{b} and Tetsuro ENDO\textsuperscript{c}

\textsuperscript{a} Department of Integrated Bioscience and Technology, Nagaoka University of Technology 1603-1 Kamitomioka, Nagaoka, Niigata 940-2188, Japan
\textsuperscript{b} Organization for the Strategic Coordination of Research and Intellectual Properties, Meiji University Kawasaki, Kanagawa 214-8571, Japan
\textsuperscript{c} Department of Electronics and Bioinformatics, School of Science and Technology, Meiji University Kawasaki, Kanagawa 214-8571, Japan

Email: qttruong@stn.nagaokaut.ac.jp, tsubone@vos.nagaokaut.ac.jp, naohiko@yomogi.jp, endoh@isc.meiji.ac.jp

Abstract—We consider chaos and bifurcation phenomena near the Chenciner bubbles generated by two coupled piecewise-constant oscillators (PWCOs) driven by a rectangular wave force. Because vector fields of PWCOs are piecewise-constants, it is relatively easy to perform rigorous analysis. For instance, a calculation algorithm for conducting Lyapunov exponent in autonomous piecewise-constant system has been proposed. By using extension of the algorithm for non-autonomous system, two-parameter Lyapunov diagrams are conducted. These results confirmed that Chaos is observed in the neighborhood of Chenciner bubbles. Furthermore, we observed the Farey sequence in the experimental measurements.

1. Introduction

Quasi-periodic bifurcations of high-dimensional tori have attracted intensive research interest in recent years. For instance, to observe bifurcation structures in discrete-time dynamics, Sekikawa et al. analyzed a coupled delayed logistic map that generated complicated quasi-periodic bifurcations [1]. Infinitely many of 2-torus generating regions existed in a 3-torus generating region, wherein 2-torus generating regions extend in numerous directions like a “cobweb” in parameter space called Arnold resonance web. The generation pattern of 2-torus regions is explained by the Farey sequence [2]. In addition, periodic solutions emerge at the intersections of two different 2-torus, which called Chenciner bubbles. The study of Arnold resonance webs is usually performed by Lyapunov analysis, and has been rapidly progressings. However, the main concerns for observing Arnold resonance webs in continuous-time dynamics are the precision and computational cost. To analyze bifurcation structures in continuous-time dynamics more precisely, Tsubone et al. proposed and analyzed piecewise-constant oscillator (PWCOs) [3]. Because vector fields of PWCOs take only constant values piecewisely, it is relatively easy to perform rigorous analysis. Hence, using the same computational cost similar to that of discrete-time dynamics, Inaba et al. succeed in observing Arnold resonance webs with high resolution in a driven continuous-time electric circuit of which governing equation is represented by two coupled piecewise-constant oscillators (PWCOs) driven by a rectangular wave force [4]. However, they obtained the Lyapunov exponents by using an ad-hoc manner. It could be difficult to apply this procedure in higher dimensional piecewise-constant dynamics. In this study, we focus on chaos and bifurcation phenomena near the Chenciner bubbles generated in a driven continuous-time electric circuit presented in [4]. The analysis is performed by using a calculation algorithm for the rigorous solutions in non-autonomous piecewise-constant system, which extended the algorithm in autonomous piecewise-constant system [5]. By using this algorithm, two-parameter Lyapunov diagrams are easily conducted. According to the numerical experiment, Chaos is observed in the neighborhood of Chenciner bubbles. Moreover, the Farey sequence is verified by laboratory measurements.

2. Piecewise-constant oscillator

In this study, we consider the piecewise-constant circuit in Fig. 1. This circuit comprises two piecewise-constant hysteresis oscillators and a rectangular wave current source. The voltage across two capacitors $C_1$ and $C_2$ is $v_1$, $v_2$, respectively. $H_1(v_1)$ and $H_2(v_2)$ are the two hysteresis elements. The waveform of the rectangular wave source with amplitude $I$ and period $2T_0$...
is shown in Fig. 2. From Kirchhoff’s law, the governing equation is represented as follows.

\[
\frac{C_1 C_2 + C_1 C_3 + C_2 C_3}{\text{d}v_1}{\text{d}t} = C_2 + C_2 H_1(v_1) + H_2(v_2) + I_1(t),
\]

\[
\frac{C_1 C_2 + C_1 C_3 + C_2 C_3}{\text{d}v_2}{\text{d}t} = C_1 + C_2 \left( H_2(v_2) + I_1(t) \right) + H_1(v_1).
\] (1)

Via rescaling,

\[
v_1 = V_{th1} x, \quad v_2 = V_{th2} y, \quad t = \gamma \tau,
\]

\[
h_1(x) h_1 = H_1(V_{th1} x), \quad h_2(y) h_2 = H_2(V_{th2} y),
\]

\[
\gamma I_{th1} C_4 \quad \gamma I_{th2} C_3 = 1, \quad I_{th1} = D_1, \quad I_{th2} = D_2, \quad I_{th1} = B,
\]

\[
C_2 + C_3 = D_3, \quad C_1 + C_3 = D_4, \quad T_0 = T,
\] (2)

the normalized equation is:

\[
\dot{x} = D_3 h_1(x) + D_2 h_2(y) + S(\tau),
\]

\[
\dot{y} = D_2 h_1(x) + D_2 D_4(D_1 h_2(y) + S(\tau)).
\] (3)

where \( h_1(x) \) and \( h_2(y) \) are the normalized hysteresis loops, of which characteristics are shown in Fig. 3(a) and Fig. 3(b), respectively. The solution on the upper branch \( h_1(x) = 1 \) jumps to the lower branch \( h_1(x) = -1 \) when \( x \) increases and reaches the point \( S_1 \). In addition, the solution on the lower branch \( h_1(x) = -1 \) jumps to the upper branch \( h_1(x) = 1 \) when \( x \) decreases and reaches the point \( S_2 \). The solution on the branch \( h_2(y) \) behaves in the same manner. Periodic external force \( S(\tau) \) is expressed as follow.

\[
S(\tau) = \begin{cases} 
B \text{ for } nT \leq \tau < (n + 1)T, \\
-B \text{ for } (n + 1)T \leq \tau < (n + 2)T.
\end{cases} \] (4)

where \( n \) is integer. \( B \) and \( T \) is the amplitude and half-periodic of the rectangular wave, respectively. The circuit dynamics include six parameters \( D_1, D_2, D_3, D_4, B, \) and \( T \). Note that because of the character of \( h_1(x) \) and \( h_2(y) \), throughout this study, we restrict our attention to the case when \( -1 \leq x(\tau) \leq 1 \) and \(-1 \leq y(\tau) \leq 1 \) hold for \( \forall \tau \). In our assumption, the trajectory in vector fields will hit boundary lines, i.e. \( x = 1, x = -1, y = 1, \) or \( y = -1 \).

3. Derivation of Lyapunov exponents in a driven piecewise-constant oscillator

In this section, we explain the procedure for deriving the Lyapunov exponents by introducing the explicit expression of the solution. In our previous work, the basic algorithm to calculate Lyapunov exponents in autonomous piecewise-constant system is proposed [5]. However, the algorithm is not suitable for the non-autonomous system, because it can not manage enforced switching depending on external force. To make it easy to conduct rigorous solutions in non-autonomous system, we assume \( \tau \) in Eq. (3) as a variable. Hence, we can rewritten Eq. (3) in the autonomous form as follows.

\[
\dot{x} = k_x D_3 h(x) + D_1 h(y) + S(z)
\]

\[
\dot{y} = k_y D_2 h(x) + D_2 D_4(D_1 h(y) + S(z))
\]

\[
\dot{z} = 1.
\] (5)

We consider the solution where the initial condition at \( z_0 \) is \( x_0 = (x_0, y_0, z_0)^\top \). When the trajectory started at \( z_0 \) hits a boundary line, i.e., either one of \( x = 1, x = -1, y = 1, y = -1, z = T, \) or \( z = 2T \) at the time \( z_1 \), the solution of Eq. (5) is expressed as follows.

\[
x_1 = x_0 + k(z_1 - z_0),
\] (6)

where \( k = (k_x, k_y, 1)^\top \) is vector fields presented by Table 1. To conduct the Jacobian matrix in a systematic manner, a normal vector \( n^\top \) is introduced. By using this normal vector, the \( (z_1 - z_0) \) in Eq. (6) is obtained.

\[
z_1 - z_0 = \frac{n^\top x_1 - n^\top x_0}{n^\top k}
\] (7)

Note that, at the time \( z_1 \) the trajectory is on boundary line. Thus, \( n^\top x_1 = D \) is a constant value presented in Table 2. For example, when the solution hits \( x = 1 \), because the normal vector is \( n = (1 0 0)^\top \), the scalar \( D \) is \( D = 1 \). In addition, substituting Eq. (7) into Eq. (6) yields the following equation.

\[
x_1 = \left( I - \frac{k n^\top}{n^\top k} \right) x_0 + \frac{k D}{n^\top k}.
\] (8)
Table 1: Values of $k$

<table>
<thead>
<tr>
<th>$k$</th>
<th>Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(D_3 + D_1 + B, D_2 + D_2 D_4(D_1 + B), 1)^\dagger$</td>
<td>$h_1(x) = 1, h_2(y) = 1, 0 \leq z &lt; T$</td>
</tr>
<tr>
<td>$(D_3 - D_3 + D_1 + B, -D_2 + D_2 D_4(D_1 + B), 1)^\dagger$</td>
<td>$h_1(x) = -1, h_2(y) = 1, 0 \leq z &lt; T$</td>
</tr>
<tr>
<td>$(D_3 - D_3 + D_1 + B, D_2 + D_2 D_4(-D_1 + B), 1)^\dagger$</td>
<td>$h_1(x) = 1, h_2(y) = -1, 0 \leq z &lt; T$</td>
</tr>
<tr>
<td>$(D_3 - D_3 - D_1 + B, D_2 + D_2 D_4(D_1 - B), 1)^\dagger$</td>
<td>$h_1(x) = -1, h_2(y) = 1, T \leq z &lt; 2T$</td>
</tr>
<tr>
<td>$(D_3 - D_3 - D_1 + B, -D_2 + D_2 D_4(-D_1 - B), 1)^\dagger$</td>
<td>$h_1(x) = 1, h_2(y) = -1, T \leq z &lt; 2T$</td>
</tr>
</tbody>
</table>

It is clear from Eq. (8) that the local Jacobian matrix $A$ is represented by

$$A = \frac{d\mathbf{x}}{d\mathbf{x}_0} = I - \frac{\mathbf{k} \mathbf{n}^\top \mathbf{n} \mathbf{k}}{\mathbf{n}^\top \mathbf{k}}.$$  \hspace{1cm} (9)

Then, if the solution hits $x = 1$ or $x = -1$,

$$A_0 = \begin{pmatrix} 0 & 0 & 0 \\ -k_y/k_x & 1 & 0 \\ -1/k_x & 0 & 1 \end{pmatrix},$$ \hspace{1cm} (10)

if the solution hits $y = 1$ or $y = -1$,

$$A_1 = \begin{pmatrix} 1 & -k_x/k_y & 0 \\ 0 & 0 & 0 \\ 0 & -1/k_y & 1 \end{pmatrix},$$ \hspace{1cm} (11)

and, if the solution hits $z = T$ or $z = 2T$,

$$A_2 = \begin{pmatrix} 1 & 0 & -k_x \\ 0 & 1 & -k_y \\ 0 & 0 & 0 \end{pmatrix}.$$ \hspace{1cm} (12)

Note that the Jacobian matrices $A_0, A_1, A_2$ includes an all zero row. Therefore, one of eigenvalues of these matrices is zero and the corresponding minimum Lyapunov exponent is $-\infty$. Hence, we define the first and the second Lyapunov exponent as follows.

$$\lambda_1 \simeq \frac{1}{N} \sum_{j=M+1}^{M+N} \ln |A^j_{1}e_1^j|$$

$$\lambda_1 + \lambda_2 \simeq \frac{1}{N} \sum_{j=M+1}^{M+N} \ln |A^j_{1}e_1^j \times A^j_{2}e_2^j|.$$ \hspace{1cm} (13)

where $e_1^j$ and $e_2^j$ are orthonormal bases. $A^j_j = \frac{d\mathbf{x}_j}{d\mathbf{x}_{j-1}}$ is the Jacobian matrix, which is one of $A_0, A_1,$ and $A_2$. $M$ and $N$ are integers. We use a sufficiently large $N$ after removing the transient state. Because the precision and speed of calculation of piecewise-constant oscillator is excellent, it is desirable to choose $M$ as large as $N$. It is reasonable to argue that for $M = N = 2 \times 10^5$ the two Lyapunov exponents converge to zero. Therefore, we consider as $\lambda_i = 0$ if the calculated $\lambda_i$ satisfies $|\lambda_i| < 1/10^6$.

4. Chenciner bubbles and Chaos generated in a piecewise-constant circuit

In this section, we fix the coupling parameters $C_3/C_1 = C_3/C_2 = 0.01$ that correspond to $D_3 = D_4 = 101$. We choose $D_1$ and $T$ as variables and set the parameters $D_2 = 1.1$ and $B = 0.005$. Fig. 4 shows two-parameter Lyapunov diagram. In this figure, a region generating a periodic solution marked in red where the oscillation frequencies of the two hysteresis oscillators and the forcing term are synchronized. This region can be denoted as the Chenciner bubbles. Regions generating two-dimensional tori ($\lambda_1 = 0, \lambda_2 < 0$) and three-dimensional tori ($\lambda_1 = 0, \lambda_2 = 0$) are marked in blue and yellow, respectively. Regions generating chaos are marked in black. As seen in Fig. 4 (a), the Arnold resonance web is clearly observed. Moreover, chaos is inevitably observed near the periodic solution generating region around which the regions generating three-dimensional tori emanate. Our numerical result shows against the remark given by Baesens et al. [6] that if the coupling parameter of dynamics equation is small, chaos generations cannot be observed.

To observing Farey sequence in the circuit, we set $C_3/C_1 = C_3/C_2 = 1/10$, i.e., $D_3 = D_4 = 11$. The nu-

<table>
<thead>
<tr>
<th>$n^+$</th>
<th>$D$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 1$</td>
<td>$(1 \ 0 \ 0)$ 1</td>
</tr>
<tr>
<td>$x = -1$</td>
<td>$(1 \ 0 \ 0)$ -1</td>
</tr>
<tr>
<td>$y = 1$</td>
<td>$(0 \ 1 \ 0)$ 1</td>
</tr>
<tr>
<td>$y = -1$</td>
<td>$(0 \ 1 \ 0)$ -1</td>
</tr>
<tr>
<td>$z = T$</td>
<td>$(0 \ 0 \ 1)$ T</td>
</tr>
<tr>
<td>$z = 2T$</td>
<td>$(0 \ 0 \ 1)$ 2T</td>
</tr>
</tbody>
</table>
Numerically obtained two-parameter diagram is shown in Fig. 4 (b). In this case, the regions generating two-dimensional tori are relatively thin. Fig. 5 (a.1), (a.2), (a.3) show the two-dimensional tori on the section \( \Pi \) which are obtained at the parameter values denoted by \( Q_1, Q_2 \) and \( Q_3 \) in Fig. 4 (b), respectively. Section \( \Pi \) obtain \( x, y \) values every periodic of rectangular wave force. The number of times the two torus touches the left boundary is denoted as \( L \). In addition, the number of times it touches the upper boundary is denoted as \( U \). In Fig. 5 (a.1), \( L = 1 \) and \( U = 0 \), and in Fig. 5 (a.3), \( L = 2 \) and \( U = 1 \). Between the parameter values at which Fig. 5 (a.1) and Fig. 5 (a.3) are obtained on the two-parameter Lyapunov diagram, the two-dimensional torus attractor with \( L = 1 + 2 = 3 \) and \( U = 0 + 1 = 1 \) exists as shown in Fig. 5 (a.2). Fig. 5(b.1), (b.2), and (b.3) show the associated experimental results. Remarkable agreement can be confirmed, and Farey sequence is observed.

5. Conclusion

We discussed Chaos and bifurcation phenomena in a driven piecewise-constant oscillator. We used a calculation algorithm for the rigorous solutions in non-autonomous piecewise-constant system to conduct Lyapunov analysis. The explicit representation of the procedure can be applicable to a wide class of piecewise-constant driven circuit. Chaos is observed in the neighborhood of Chenciner bubbles. Moreover, the Farey sequence is observed in the experimental measurements. In future work, we plan to discuss quasiperiodic bifurcation in higher dimensional piecewise-constant oscillator.

References

Homoclinic bifurcations in a piece-wise constant neuron model

Chiaki Matsuda† and Hiroyuki Torikai‡
†Department of Computer Science, Kyoto Sangyo University
‡Department of Information Engineering, Kyoto Sangyo University
Motoyama, Kamigamo, Kita-Ku, Kyoto 603-8555, Japan

g1345297@cse.kyoto-su.ac.jp, torikai@cse.kyoto-su.ac.jp

Abstract—In this paper, a modified piece-wise constant neuron model is presented. It is shown that the circuit exhibits a homoclinic bifurcation known as a blue-sky catastrophe, which is also observed in a conductance-based neuron model. Also, a bifurcation diagram is obtained based on numerical experiments. Using the bifurcation diagram, occurrence mechanism of the bifurcation is explained.

1. Introduction
A piece-wise constant (PWC) neuron model has a piece-wise constant vector field and has been designed to mimic bifurcations of neurons [1]-[3]. Advantages of the PWC neuron model include the following points: (a) it can be implemented by using standard circuit elements (highly specialized device technology is not needed), (b) its dynamics can be analyzed by using bifurcation analysis techniques for piece-wise constant vector fields, and (c) its vector field is relatively robust against parameter mismatch. It has been shown that the PWC neuron model can exhibit bifurcations of resting states (stable equilibrium points) such as supercritical and subcritical Hopf bifurcations, saddle-node on and off invariant circle bifurcations, and saddle homoclinic bifurcation [1]-[3], where these bifurcations have been analyzed extensively. It should be emphasized that these bifurcations are underlying mechanisms of neuron-like nonlinear responses of the PWC neuron model such as class 1 excitability, class 2 excitability, IF-curve with hysteresis, and IF-curve without hysteresis. In addition, recently, it was shown that the PWC neuron model can also exhibit bifurcations of spiking states (stable periodic orbits) such as a homoclinic bifurcation of spiking states known as a blue-sky catastrophe [1]-[3]. However, analysis of the homoclinic bifurcation of the PWC neuron model has been insufficient so far.

In this paper, a modified PWC neuron model is presented. It is shown that the modification leads to occurrence of a blue-sky catastrophe, which is more similar to that of a Hodgkin-Huxley type conductance-based neuron model [4] than our previous PWC neuron model [3]. Then, a bifurcation diagram is obtained based on numerical simulations. Using the bifurcation diagram, occurrence mechanism of the blue-sky catastrophe is explained. It should be emphasized that the analysis result in this paper will be a preliminary result to develop theoretical analysis methods of homoclinic bifurcations of the PWC neuron model and to develop bifurcation-based design methods of the PWC neuron model.

2. Blue-sky catastrophe
Let us begin with a Hodgkin-Huxley type conductance-based neuron model is described by the following equation [5]:

\[
\frac{dV}{dt} = g_{K}^{2}(V - E_{K}) - g_{L}^{2}(V - E_{L}) - g_{Na}^{2}(V - E_{Na})(V - E_{Na}) + \mu V,
\]

\[
\frac{dI_{pol}}{dt} = f(-83.018 + V^{2} - 2y),
\]

where \(V\) is a membrane potential and \(m_{Na}, h_{Na}\) are gate variables. Also, \(f\) is a nonlinear function given by \(f(a, b, V) = (1 + e^{a(V+b)})^{-1}\). It is shown that the above model exhibits a homoclinic bifurcation known as the blue-sky catastrophe. In order to investigate the essential structure of the blue-sky catastrophe, the following normal form is sometimes used [4].

\[
dw/dt = x(2 + \mu - b(x^{2} + y^{2})) + z^{2} + y^{2} + 2y,
\]

\[
dy/dt = -z^{2} - (y + 1)(x^{2} + y^{2} + 2y) - 4x + \mu y,
\]

\[
dz/dt = z^{2}(y + 1) + x^{2} - \epsilon.
\]

Fig. 1 shows time waveforms of the normal form of the blue-sky catastrophe. Fig. 1(a), the variable \(v\) is oscillating with very high frequency. Fig. 1(b), the parameter \(\mu\) is slightly changed and the variable \(x\) suddenly starts to burst. Fig. 1(c), the parameter \(\mu\) is further changed and the interburst-interval of the variable \(x\) becomes shorter. Since the complicated bursting orbit (catastrophe) in Fig. 1(b) seems to appear suddenly from somewhere (blue-sky), which is not related to the periodic orbit in Fig. 1(a), this bifurcation phenomenon is called a blue-sky catastrophe.

3. Modified PWC Neuron Model
Fig. 2(a) shows a modified piece-wise constant (PWC) neuron model. The capacitor voltages \(v\) and \(u\) correspond to a membrane potential and recovery variable of a neuron model, respectively. As shown in Fig. 2(b)-(d), the current sources \(I_{e}, I_{u}\), and \(I_{M}\) are voltage-controlled and have the following PWC characteristics, i.e., step-function-like characteristics.

\[
I_{e}(v_{e}) = \begin{cases} +I_{e}^{+} & \text{if } v_{e} \geq 0, \\ -I_{e}^{-} & \text{if } v_{e} < 0, \end{cases}
\]

\[
I_{u}(u_{e}) = \begin{cases} +I_{u}^{+} & \text{if } u_{e} \geq 0, \\ -I_{u}^{-} & \text{if } u_{e} < 0, \end{cases}
\]
Due to the PWC characteristics of the current sources, the PWC neuron model has a piece-wise constant vector field. The switch is voltage-controlled and realizes a firing reset of the membrane potential \( v \) to a voltage-controlled reset level \( V_B \), which has the following characteristics as shown in Fig. 1(e).

\[
I_{\text{MH}'}(v, u) = \begin{cases} 
-I_{\text{MH}} & \text{if } v \geq V_B(u) \\
+I_{\text{MH}}' & \text{otherwise.}
\end{cases}
\]

Then, the dynamics of the PWC neuron model is described by the following equations.

If \( v(t) < V_A \) (i.e., the switch \( SW \) is opened), then

\[
\frac{dv}{dt} = I_v(|v| + V_{in} - u),
\]

\[
\frac{du}{dt} = I_u(a v - u) + I_{\text{MH}}(v, u).
\]

If \( v(t) = V_A \) (i.e., the switch \( SW \) is closed), then

\[
v(t^*) = V_B(t).
\]

where \( t^* = \lim_{\delta \to 0} t + \delta, \delta > 0 \). Fig. 3 shows typical time waveforms of the membrane potential \( v \) and the recovery variable \( u \) of the PWC neuron model. Fig. 4 shows a vector field of the PWC neuron model. As shown in Fig. 4, the characteristics of the voltage controlled voltage source \( V_B \) is controlled by the recovery variable \( u \). Due to this voltage controlled voltage source \( V_B \), the time waveform of the membrane potential \( v \) becomes more similar to that of the conductance-based neuron model compared to our previous model [3].
Figure 3: Typical time waveforms of the (a) membrane potential $v$ and (b) the recovery variable $u$ of the PWC neuron model.

4. Analysis of Homoclinic Bifurcation

Fig. 5 shows typical time waveforms and corresponding phase plane trajectories of the PWC neuron model. In Figs. 5(a) and (a'), the PWC neuron model has a stable periodic orbit (SB) corresponding to a spiking orbit with high frequency and an unstable periodic orbit (UB), where UB is shown by a dashed orbit in (a') and is not shown in (a). In Figs. 5(b) and (b'), the parameter $I_{MH}$ is slightly changed from (a) and (a'). In this case, the SB and UB merged and disappeared. Instead of the disappeared orbits, a complicated orbit suddenly appeared as shown in 5(b) and (b'). In Figs. 5(c) and (c'), the parameter $I_{MH}$ is further changed from (b) and (b'). In this case, the inter-burst-interval of the membrane potential $v$ becomes shorter. The mechanism of the above change of phenomena is conceptually same as the occurrence mechanism of the blue-sky catastrophe [4]. In order to analyze the blue-sky catastrophe, a two-parameter bifurcation diagram in Fig. 6 is obtained by numerical experiments. The points (a)-(c) in the diagram correspond to Figs. 5(a) and (a')-5(c) and (c'), respectively. In the region (A), the PWC neuron model has an SB and a UB. In the region (B), the PWC neuron model has a complicated bursting orbit. On the border of the regions (A) and (B), the SB and the UB merge and disappear and the complicated bursting orbit appears. At the point (c), properties of unstable manifold changes from those in (A) and (B). Note that this bifurcation diagram will be a preliminary ingredient to develop a systematic design procedure of the PWC neuron model to exhibit homoclinic bifurcations.

5. Conclusions

In this paper, the modified PWC neuron model was investigated. It was shown that the model exhibits the blue-sky catastrophe, which is also observed in a conductance-based neuron model. Also, a bifurcation diagram was obtained based on the numerical experiments. Using the diagram, the occurrence mechanism of the blue-sky catastrophe was explained. Future problems include: (a) theoretical analysis of the blue-sky catastrophe and other homoclinic bifurcations in the PWC neuron model, (b) development of a network of the PWC neuron models, and (c) their CMOS implementations. This work was partially supported by JSPS KAKENHI Grant Number 15K00352.

References


Figure 5: Bifurcation of homoclinic periodic orbits. (a)-(c) show time waveforms and (a')-(c') show corresponding phase plane orbits. SB and UB stand for "stable periodic orbit" and "unstable periodic orbit." $V_B=0.6$, $V_A=1.0$, $I^+_u=0.019$, $I^-_u=0.33$, $I^+_v=0.16$, $I^-_v=0.135$, $a=10.0$, $V_m=0.6$. (a) and (b) Coexistence of SB and UB. $I^+_{MH}=0.07$, $I^-_{MH}=0$, $U^+_{MH}=0.85$, $U^-_{MH}=0.6$. (b) and (b') Disappearance of the SB and UB and appearance of a complicated bursting orbit. $I^+_{MH}=0.015$ and $U^-_{MH}=0.6$. (c) and (c') Another disappearance of the SB and UB and appearance of a complicated bursting orbit. $I^+_{MH}=0.07$ and $U^-_{MH}=0.85$.

Figure 6: Bifurcation diagrams obtained by numerical experiments. The points (a)-(c) in the diagram correspond to Figs. 5(a) and 5(a')-5(c) and (c'), respectively. In region (A), the PWC neuron model has an SB and a UB. In region (B), the PWC neuron model has a complicated bursting orbit. On the border of the regions (A) and (B), the SB and the UB merge and disappear and the complicated bursting orbit appears.

- 159 -
Qualitative Behavior of Nonideal Switching Circuit with Two-periodic Inputs

Sho Hayashida†, Hiroyuki Asahara‡, Takuji Kousaka†

†Faculty of Engineering, Oita University, 700 Dannoharu, Oita-shi, Oita 870-1192, Japan.
‡Faculty of Engineering, Okayama University of Science, 1-1 Ridai-cho, Kita-ku Okayama-shi, Okayama 700-0005, Japan.
Email: hayashida@bifurcation.jp, asahara@ee.ous.ac.jp, takuji@oita-u.ac.jp

Abstract—This paper studies the dynamical effect of spike noise in a switching circuit with two-periodic inputs. We explain behavior of the circuit. Then, we consider the dynamical effect of spike noise by comparing the system with ideal switching and the system with nonideal switching.

1. Introduction

Power conversion circuit, such as a converter circuits and inverter circuits are the typical example of the interrupted electric circuit. Nonlinear phenomena in these circuits are analyzed from nonlinear dynamics in the interrupted electric circuit from numerical and experimental viewpoints [1]. Interrupted dynamical systems are analyzed under the assumption of ideal switching. On the other hand, it was reported that Banerjee et al reported that the spike noise arises immediately after the switching action and greatly influences the bifurcation structure of the systems [2].

In this paper, we examine the dynamical effect of spike noise in a interrupted electric circuit with tow-periodic inputs.

2. Interrupted electric circuit and spike noise

We consider the following two differential systems.

\[
\frac{dy}{d\tau} = \begin{cases} 
-y + B, & \text{system-a} \\
-y + A\sin\Omega\tau, & \text{system-b}
\end{cases}
\] (1)

The system has two different periodic external forces; the clock pulse and the sinusoidal signal. \(T_f\) and \(T_s = NT_f\) denote the period of the clock pulse and that of the sinusoidal signal, respectively. Figure 1 shows the examples of the waveform with non ideal switching. We assume that spike noise arise when system is replaced.

Now, The system-a changes to system-b when the waveform reaches the reference value \(y_r\). Following, the system-b changes to system-a when the next clock pulse arrives. Note that switching action from system-b to system-a does not occur if the spike noise reaches to reference value as shown in gray colored area in Fig. 1. Here, We use the parameter: \(N = 30, A = 0.1, B = 1.0, T_s = 0.8, h = 0.1\).

3. Effect of spike noise

Figure 2 shows the 1-parameter bifurcation diagram. The new bifurcation phenomena arises when the spike noise occurs. We know that the new types of the orbit behavior which is caused by the spike noise makes the coexistence region.

4. Conclusion

This paper studied the dynamical effect of spike noise by comparing the system with ideal switching and the system with nonideal switching.

References


Modeling and Simulation of Motion of an Underwater Robot

Ryo Inohara†, Kaito Isogai†, Hideo Nakano‡, and Hideaki Okazaki†

†Graduate School of Engineering, Shonan Institute of Technology
1–1–25, Tsujidounishikaigan, Fujisawa-shi, Kanagawa Prefecture 251-8511, Japan
‡Faculty of Engineering, Shonan Institute of Technology
1–1–25, Tsujidounishikaigan, Fujisawa-shi, Kanagawa Prefecture 251-8511, Japan
Email: 15t2002@sit.shonan-it.ac.jp, 15t2001@sit.shonan-it.ac.jp, nakano@sc.shonan-it.ac.jp, okazaki@sc.shonan-it.ac.jp

Abstract—This paper presents how system dynamics and system control equations for an underwater robot were derived using an Arnold-type operator to control the OpenROV. Typical behavior of the OpenROV on MATLAB numerical simulations is illustrated.

1. Introduction

Although there are several designs, control system equations, and dynamic equations for underwater robots, such as [1], unified methods to describe the dynamic equations for the rigid body kinetics of an underwater robot have yet to be established. Using the mathematical foundation of rigid body dynamics provided by V. I. Arnold [2], we apply Arnold’s operator in order to facilitate the derivation of dynamics for the rigid body kinetics of an underwater robot. Secondly, OpenROV (Fig 1) (open-source remotely operated vehicle) projects [3] have recently been promoted to examine the sea bottom. Since simulations of the behavior of OpenROV are of value, we describe the equations of motion for real OpenROV. Finally, we illustrate the typical behavior of OpenROV on MATLAB numerical simulations. In the following, let \( \mathbb{R} \) be the set of real numbers and \( \mathbb{R}^n \) be the set of real number vectors.

\[ B : W \to w, \]

which preserves the metric and the orientation (Fig 2).

2. Motion in a Moving Coordinate System

In this section, we detail the mathematical foundation for describing the motion of fundamental rigid body kinematics [1] of an underwater robot based on [2], [4]. The time parameter \( t \) for all the stated variables, such as \( r(t) \), or \( \Omega(t) \), etc., is omitted for convenience. We use the following notation as [2] (Fig 2):

\[ e_i \in w \quad (i = 1, 2, 3) \]

are the base vectors of a right-handed Cartesian stationary coordinate system at the origin \( O \);

\[ E_i \in W \quad (i = 1, 2, 3) \]

are the base vectors of a right moving coordinate system connected to the body at the center of the mass \( O_c \).

**Definition 1** Let \( w \) and \( W \) be oriented euclidean spaces (i.e. orthogonal spaces). A motion of \( W \) relative to \( w \) is a smooth mapping on \( t \):

\[ B : W \to w, \]

**Definition 2** A motion \( B \) is called a rotation if it takes the origin of \( W \) to the origin of \( w \) (i.e. if \( B \) is a linear operator).

**Definition 3** \( w \) is called a stationary coordinate system, \( W \) a moving one, and \( q \in w \) the radius-vector of a point moving relative to the stationary system; if

\[ q = r + ut + BQ \]

There exists a flow of fluid velocity vector \( u \) in \( w \). \( Q \) is called the radius vector of the point relative to the moving system (Fig 2). We express the “absolute velocity” \( \dot{q} \) in
terms of the relative motion $Q$ and the motion of the coordinate system $B$. By differentiating with respect to $t$ in Eq. (2), we arrive at Eq. (3) for the addition of velocities.

$$\dot{q} = \dot{r} + \dot{u} + u = \frac{d}{dt}(BQ).$$

(3)

In order to carry the stationary frame $e_i$ ($i = 1, 2, 3$) into the moving frame $E_i$ ($i = 1, 2, 3$), we perform three rotations (Fig 3):

1. Given an angle $\psi$ around the $e_3$ axis, under this rotation, $e_3$ remains fixed and $e_2$ goes to $E_2^{-2}$ by means of Eq. (5).

2. Given an angle $\theta$ around the $E_2^{-2}$ axis, under this rotation, $E_2^{-2}$ remains fixed and $E_2^{-1}$ goes to $E_1^{-1}$ by means of Eq. (6).

3. Given an angle $\phi$ around the $E_1^{-1}$ axis, under this rotation, $E_1^{-1}$ remains fixed and $E_3$ goes to $E_3$ by means of Eq. (7).

After all three rotations are completed, $e_1$ has moved to $E_1$, and $e_2$ to $E_2$; therefore, $e_1$ moves to $E_3$. The angles $\psi$, $\theta$, and $\phi$ are called the Tait-Bryan angles (one of the Euler angle systems).

![Fig 3: Rotations defining the Tait-Bryan angles](image)

Here we describe an operator $B$, as follows:

$$B = R_{\psi} R_{\theta} R_{\phi},$$

$$= \begin{pmatrix}
\cos \phi & -\sin \phi & 0 \\
\sin \phi & \cos \phi & 0 \\
0 & 0 & 1
\end{pmatrix},$$

(4)

$$R_{\psi} = \begin{pmatrix}
\cos \psi & -\sin \psi & 0 \\
\sin \psi & \cos \psi & 0 \\
0 & 0 & 1
\end{pmatrix},$$

(5)

$$R_{\theta} = \begin{pmatrix}
\cos \theta & 0 & \sin \theta \\
0 & 1 & 0 \\
-\sin \theta & 0 & \cos \theta
\end{pmatrix},$$

(6)

$$R_{\phi} = \begin{pmatrix}
1 & 0 & 0 \\
0 & \cos \phi & -\sin \phi \\
0 & \sin \phi & \cos \phi
\end{pmatrix}.$$  

(7)

c$\psi$, $s\psi$, $c\theta$, $s\theta$, $c\phi$ and $s\phi$ denote $\cos \psi$, $\sin \psi$, $\cos \theta$, $\sin \theta$, $\cos \phi$ and $\sin \phi$, respectively. Hence, the base vectors of the moving coordinate system generated by means of the operator $B$ are expressed as:

$$BE_1 = \cos \psi \cos \theta e_1 + \sin \psi \cos \theta e_2 - \sin \theta e_3,$$

(8)

$$BE_2 = (\cos \psi \sin \theta \sin \phi - \sin \psi \cos \phi) e_1 + \sin \psi \sin \theta \cos \phi + \cos \psi \cos \phi e_2$$

+ $\cos \theta \sin \phi e_3,$

(9)

$$BE_3 = (\cos \psi \sin \theta \cos \phi + \sin \psi \sin \phi) e_1$$

+ $(\sin \psi \sin \theta \cos \phi - \cos \psi \sin \phi) e_2$

+ $\cos \theta \cos \phi e_3.$

(10)

Since $W$ is a moving coordinate system connected to the body of an underwater robot, $Q$ is at rest in $W$ (i.e., $\dot{Q} = 0$) and the coordinate system $W$ rotates (i.e., $r = 0$). In this case, the motion of the point $q$ is a transferred rotation given by Eq. (11).

$$q = r + \dot{u} + u + BQ = r + \dot{u} + u + [B\Omega, BQ],$$

(11)

where $[\cdot, \cdot]$: the vector product.

The vector $\Omega \in W$ is called the vector of angular velocity in the underwater robot. In this case, $\Omega$ is expressed by:

$$\Omega = B^T \omega.$$  

(12)

The vector $\omega \in w$ is called the instantaneous angular velocity given by Eq. (13).

$$\omega = \psi e_3 + \theta E_2^{-2} + \phi E_3^{-1}.$$  

(13)

In numerous studies and texts, the angular velocity vector $(\psi, \theta, \phi)$ of the Tait-Bryan angles $[5], [6]$ is often referred to as coordinate components $E_i$, ($i = 1, 2, 3$) on the base vectors of the moving coordinate system. It must be stressed, however, that the description given in Eq. (13) is correct. Using the angular velocity vector of the Tait-Bryan angles of Eq. (13), we can rewrite Eq. (11), as follows:

$$q = r + \dot{u} + u + \psi \frac{\partial}{\partial \psi} (BQ) + \theta \frac{\partial}{\partial \theta} (BQ) + \phi \frac{\partial}{\partial \phi} (BQ).$$  

(14)

Here, let $h \in w$ be the angular momentum of the underwater robot in the stationary inertia coordinate system, $H \in W$ be the angular momentum of the underwater robot in the moving coordinate system, and $\dot{I}$ be the moment of inertia of the underwater robot. Using operator $B$, we obtained the following equations:

$$h = \dot{I} \omega = BH \in w,$$

(15)

$$H = \dot{I} \Omega \in W.$$  

(16)

In addition, let $\tau \in w$ be the torque of the underwater robot. We obtain the time derivative of an angular momentum which is equal to the moment, as follows:

$$\frac{d}{dt} h = \tau = \frac{d}{dt} BH = BT, \dot{I} \Omega + \Omega H - T = 0.$$  

(17)

Then, $\Omega$ and $\dot{\Omega}$ can be expressed in concrete terms by:

$$\Omega = (-\psi \sin \theta + \phi) E_1 + (\psi \cos \theta \sin \phi + \theta \cos \phi) E_2$$

+ $(\psi \cos \theta \cos \phi - \theta \sin \phi) E_3,$

(18)

and

$$\dot{\Omega} = (-\dot{\psi} \sin \theta - \dot{\phi} \cos \theta + \dot{\phi}) E_1$$

+ $(\dot{\psi} \cos \phi - \dot{\psi} \cos \theta \sin \phi + \psi \dot{\theta} \cos \phi + \theta \dot{\psi} \cos \phi) E_2$

+ $(\dot{\psi} \cos \phi - \dot{\phi} \sin \phi - \psi \dot{\theta} \sin \phi - \theta \dot{\psi} \sin \phi) E_3,$

(19)

respectively.
A torque $\tau$ of Eq. (17) is also given as:

$$\tau = [r, f].$$

(20)

Here, $r$ and $f$ denote the position vector on which external forces act and a vector of external forces, respectively.

The coordinate components of an angular momentum of the underwater robot are given as Eq. (21).

$$H_A = \sum_{i=1}^{3} H_{Ai} E_i.$$  

(21)

Then, we have the following equation:

$$(H_1, H_2, H_3)^T = \hat{I}(\Omega_1, \Omega_2, \Omega_3)^T.$$  

(22)

The moment of inertia $\hat{I}$ is also defined by

$$\hat{I} = \begin{bmatrix} I_{11} & I_{12} & I_{13} \\ I_{21} & I_{22} & I_{23} \\ I_{31} & I_{32} & I_{33} \end{bmatrix}.$$  

(23)

In this paper, for $\hat{I}$, we assume that $I_{33} = 0$, and for $k \neq l$.

3. Dynamic Model of OpenROV

![Figure 4: Coordinate systems and forces/moments acting on OpenROV](image)

Fig 4: Coordinate systems and forces/moments acting on OpenROV ($m = 2.4$ [kg], $g = 9.80665$ [m/s²], $I_{11} = 0.01355$ [kg · m²], $I_{22} = 0.00480$ [kg · m²], $I_{33} = 0.00593$ [kg · m²], $l = 0.165$ [m]).

The coordinate systems and free body diagram for OpenROV are shown in Figure 4. Based on the preceding mathematical foundation, we can describe the dynamic model of OpenROV (Fig. 4). $F_i, (i = 1, 2, 3)$ and $T_i, (i = 1, 2, 3)$ of Figure 4, represent vertical forces and moment, respectively. $F_i, (i = 1, 2, 3)$ and $T_i, (i = 1, 2, 3)$ are defined in a similar manner [7]. Each motor of OpenROV has an angular speed $\omega_i$ and produces a vertical force $F_i$ according to:

$$F_i = k_F \omega_i^2,$$  

(24)

Experimentation with a fixed motor in a steady state shows that $k_F = 7.3 \times 10^{-6}$ [Nm/rpm]. The motors also produce a moment according to:

$$T_i = k_T \omega_i^2,$$  

(25)

The constant, $k_T$, is determined to be approximately $1.5 \approx 2 \times 10^{-5}$ [Nm/rpm] by matching the performance of the simulation to the real system. For the rigid body kinetics of Eq.(3) and Eq.(17), taking into consideration gravitational or buoyancy terms, system inertia matrix (including added mass), system moment matrix (including added moment), viscous damping, current loads, and an assumption of $\dot{u} = 0$, we arrive at the equations of motion for OpenROV, as follows:

$$(M + M_{add})(\dot{r} + 2\dot{u}) = (\rho V_{vol} - m)g \mathbf{e}_3 - k_D |\dot{r} - u| (\dot{r} - u)$$

$$+ k_L |\dot{r} - u| \left( \sum_{i=1}^{3} B E_i - (r - u, B E_i) \right)$$

$$- \Gamma_i (\dot{r} - u) + B \left( -F_1 E_1 - F_2 E_2 + F_3 E_3 \right),$$  

(26)

$$\hat{I} \hat{\dot{\mathbf{\omega}}} + [\mathbf{\Omega}, \hat{I} \hat{\dot{\mathbf{\omega}}} + \hat{\mathbf{\Omega}}] \hat{\mathbf{\Omega}} = -\Gamma_i \hat{\mathbf{\Omega}} + (T_1 - T_2) E_i$$

$$+ \left[ E_2, F_1 E_1 \right] + \left[ -E_2, F_1 E_1 \right].$$  

(27)

$m$: the gross weight of OpenROV, $M = m \times \text{Unitmatrix}$, $\rho$: fluid density, $V_{vol}$: the volume of the fluid displaced by OpenROV, $g$: the gravitational acceleration, $\rho V_{vol}$: buoyancy, $k_D$: drift force coefficient function, $k_L$: lift force coefficient function, $\Gamma_i$: viscous damping coefficient functions, $M_{add}$: added mass matrix, $\hat{I}_{add}$: added moment of inertia, $k_D = k_D(u, r, \psi, \theta, \phi)$, and $k_L = k_L(u, r, \psi, \theta, \phi)$. Notice that $M_{add}$, $\hat{I}_{add}$, $\Gamma_i$, and $\Gamma_i$ are diagonal matrices.

Since the operator $B$ states variables $\Omega$ and $\mathbf{\Omega}$ are expressed as the functions of $(\psi, \dot{\psi}, \phi, \dot{\phi}, \theta, \dot{\theta}, \dot{\phi})$, the state equation of OpenROV can be rewritten as equations of the function of $(\psi, \dot{\psi}, \phi, \dot{\phi}, \theta, \dot{\theta}, \dot{\phi})$.

Note that OpenROV can shift the direction of rotation of its motors using $s_i, (i = 1, 2, 3)$. When OpenROV moves forward, however, Motor 1 and Motor 2 rotate at differential directions to maintain the balance of OpenROV. Furthermore, the propeller pitch of Motor 1 is different from the propeller pitch of Motor 2.

$$\dot{x} = F(x) + G(x, u),$$  

(28)

$$x^T = (\psi, \dot{\psi}, \phi, \dot{\phi}, \theta),$$  

(29)

$$\delta x^T = (\delta \psi, \delta \dot{\psi}, \delta \phi, \delta \dot{\phi}, \delta \theta),$$  

(30)

$$u^T = ((-1)^i \omega_{\phi i 1}, (-1)^i \omega_{\phi i 2}, (-1)^i \omega_{\phi i 3}),$$  

(31)

$$F(x) = \begin{bmatrix} f_\psi(\psi, \dot{\psi}, \phi, \dot{\phi}, \theta) \\ f_\phi(\psi, \dot{\psi}, \phi, \dot{\phi}, \theta) \\ f_\theta(\psi, \dot{\psi}, \phi, \dot{\phi}, \theta) \end{bmatrix},$$  

(32)

$$B_G = \begin{bmatrix} \frac{1}{(\sin \theta - 1) \Omega_{\phi i 1} \Omega_{\phi i 1}} B_{i \phi i 1}^T \\ \frac{1}{(\sin \theta - 1) \Omega_{\phi i 1} \Omega_{\phi i 1}} B_{i \phi i 2}^T \\ \frac{1}{(\sin \theta - 1) \Omega_{\phi i 1} \Omega_{\phi i 1}} B_{i \phi i 3}^T \end{bmatrix}.$$  

(33)
The variational equation (34) that can be used to control the OpenROV is described by

$$\frac{d}{dt} \delta x = DF(x_0) \cdot \delta x + B_G u,$$  

(34)

where $DF(x_0)$ and $x_0$ denote the Jacobian of $F(x)$ and the driving point of $x$. In particular, $f_x$, $f_y$, and $f_z$ in $F(x)$ are easily obtained using Maple symbolic computations on $B^T_G$, $B^T_G$, and $B^T_G$, as follows:

$$B^T_G = [\ell \cos \phi \cos \theta k_{F1}, -\ell \cos \phi \cos \theta k_{F2}, -\cos \phi \cos \theta k_{M1}],$$  

(35)

$$B^T_G = [\ell \sin \phi k_{F1}, -\ell \sin \phi k_{F2}, \sin \phi k_{M3}],$$  

(36)

$$B^T_G = [(I_{33} + I_{333}) \sin^2 \theta k_{M1} + (I_{11} + I_{333}) \ell \cos \phi \cos \theta \sin \theta k_{F1} - (I_{33} + I_{333})k_{M1},$$  

$$+ (I_{11} + I_{333}) \ell \cos \phi \cos \theta \sin \theta k_{F2} - (I_{33} + I_{333})k_{M2}, -(I_{33} + I_{333}) \cos \phi \cos \theta \sin \theta k_{M3}].$$  

(37)

4. Simulation of Motion of OpenROV

By means of numerical computations on MATLAB with ode45 solver applied to Eq. (26) and Eq. (28), the simulation results illustrated in Figure 5 and Figure 6 are obtained. The initial values are set as $\rho V_{vol} = 23.53596 [N]$, $k_D = 0.015$, $k_L = 0.5$, $\Gamma = \Gamma_r = 2$, $M_{add} = 4$, $I_{add} = 0$, $\psi = \theta = \phi = 0 [rad/s]$, $\psi = \theta = \phi = 0 [rad]$, $r_1 = r_2 = 0 [m]$, $r_3 = -1 [m]$, $v_1 = r_2 = r_3 = 0 [m/s]$, $u_1 = 0.1 [m/s]$, $u_2 = 0.2 [m/s]$, $u_3 = -0.1 [m/s]$, $\omega_{M1} = 1000 [rpm]$, $\omega_{M2} = 900 [rpm]$, $\omega_{M3} = 0 [rpm]$, and $s_1 = s_2 = s_3 = 0$.

5. Concluding remarks

The following results are obtained:

(1) We have derived the system dynamics and system control equations for an underwater robot by using the operator $B$ together with two coordinate systems, $W$ and $w$, where simple ocean currents exist. In addition, we have described the equations of motion for a real OpenROV version 2.7.

(2) We have reliably illustrated the typical behavior of the OpenROV by numerical simulations on MATLAB with an ode45 solver; however, appropriate numerical methods for reliable simulations should be investigated further.

References


Modeling and Simulation of Motion of a Quadcopter

Kaito Isogai, Ryo Inohara, Hideo Nakano, and Hideaki Okazaki

1. Introduction

Although there are several designs, control system equations, and dynamic equations for quadcopters, such as [1], unified methods to describe the dynamic equations of quadcopters have yet to be established. Using the mathematical foundation of rigid body dynamics provided by V. I. Arnold [2], we apply Arnold’s operator in order to facilitate the derivation of the dynamics of a quadcopter. Secondly, we clarify the relationship between the Euler angles, the stationary inertia coordinate system and the moving coordinate system for engineering problems. Finally, we illustrate the typical behavior of the quadcopter on MATLAB numerical simulations. In the following, let be the set of real numbers and be the set of real number vectors.

2. Motion in a Moving Coordinate System

In this section, we detail the mathematical foundation for describing the motion of a quadcopter based on [2]. It should be noted that the method applied to describe the motion in Inohara et al. [3] is used for this study. The time parameter for all the stated variables, such as or, etc., is omitted for convenience. We use the following notation as [2] (Fig 1):

\[ e_i \in W \ (i = 1, 2, 3) \] are the base vectors of a right-handed Cartesian stationary coordinate system at the origin \( O \);
\[ E_i \in W \ (i = 1, 2, 3) \] are the base vectors of a right moving coordinate system connected to the body at the center of the mass \( O_c \).

**Definition 1** Let \( W \) and \( W \) be oriented euclidean spaces (i.e. orthogonal spaces). A motion of \( W \) relative to \( W \) is a smooth mapping depending on \( t \):

\[ B : W \rightarrow W, \] which preserves the metric and the orientation (Fig 1).

**Definition 2** A motion \( B \) is called a rotation if it takes the origin of \( W \) to the origin of \( w \) (i.e. \( B \) is a linear operator).

**Definition 3** \( w \) is called a stationary coordinate system, \( W \) a moving one, and \( q \in w \) the radius-vector of a point moving relative to the stationary system; if

\[ q = r + BQ \] (2)

\( Q \) is called the radius vector of the point relative to the moving system (Fig 1). We express the “absolute velocity” \( q \) in terms of the relative motion \( Q \) and the motion of the coordinate system \( B \). By differentiating with respect to \( t \) in Eq. (2), we arrive at Eq. (3) for the addition of velocities.

\[ \dot{q} = \dot{r} + BQ + B\dot{Q}. \] (3)

In order to carry the stationary frame \( e_i \ (i = 1, 2, 3) \) into the moving frame \( E_i \ (i = 1, 2, 3) \), we perform three rotations (Fig 2):

1. Given an angle \( \psi \) around the \( e_3 \) axis, under this rotation, \( e_3 \) remains fixed and \( e_2 \) goes to \( E_2^{-2} \) by means of Eq. (5).
2. Given an angle \( \theta \) around the \( E_2^{-2} \) axis, under this rotation, \( E_2^{-2} \) remains fixed and \( E_1 \) goes to \( E_1^{-1} \) by means of Eq. (6).
3. Given an angle $\phi$ around the $E_1^{-1}$ axis, under this rotation, $E_1^{-1}$ remains fixed and $E_3$ goes to $E_3$ by means of Eq. (19).

After all three rotations are completed, $e_1$ has moved to $E_1$, and $e_2$ to $E_2$; therefore, $e_1$ moves to $E_3$. The angles $\psi$, $\theta$, and $\phi$ are called the Tait-Bryan angles (one of the Euler angle systems).

Fig 2: Rotations defining the Tait-Bryan angles

Here we describe an operator $B$, as follows:

$$B = R_\psi R_\theta R_\phi = \begin{pmatrix}
\cos \psi & -\sin \psi & 0 \\
\sin \psi & \cos \psi & 0 \\
0 & 0 & 1
\end{pmatrix},$$

$$R_\psi = \begin{pmatrix}
\cos \theta & 0 & \sin \theta \\
0 & 1 & 0 \\
-\sin \theta & 0 & \cos \theta
\end{pmatrix},$$

$$R_\phi = \begin{pmatrix}
1 & 0 & 0 \\
0 & \cos \phi & -\sin \phi \\
0 & \sin \phi & \cos \phi
\end{pmatrix}. \tag{4}$$

$c_\psi$, $s_\psi$, $c_\theta$, $s_\theta$, $c_\phi$ and $s_\phi$ denote $\cos \psi$, $\sin \psi$, $\cos \theta$, $\sin \theta$, $\cos \phi$ and $\sin \phi$, respectively. Hence, the base vectors of the moving coordinate system generated by means of the operator $B$ are expressed as:

$$BE_1 = \cos \psi \cos \theta e_1 + \sin \psi \cos \theta e_2 - \sin \theta e_3, \tag{5}$$

$$BE_2 = (\cos \psi \sin \theta \sin \phi - \sin \psi \cos \theta) e_1 + (\sin \psi \sin \theta \sin \phi + \cos \psi \cos \phi) e_2 + \cos \theta \sin \phi e_3, \tag{6}$$

$$BE_3 = (\cos \psi \sin \theta \cos \phi + \sin \psi \sin \phi) e_1 + (\sin \psi \sin \theta \cos \phi - \cos \psi \sin \phi) e_2 + \cos \theta \cos \phi e_3. \tag{7}$$

Since $W$ is a moving coordinate system connected to the body of a quadcopter, $Q$ is at rest in $W$ (i.e., $Q = 0$) and the coordinate system $W$ rotates (i.e., $r = 0$). In this case, the motion of the point $q$ is a transferred rotation given by Eq. (11).

$$q = r + BQ = r + B[\Omega, Q] = r + [B\Omega, BQ]. \tag{11}$$

where $[\cdot, \cdot]$: the vector product.

The vector $\Omega \in W$ is called the vector of angular velocity in the quadcopter. In this case, $\Omega$ is expressed by:

$$\Omega = B^T \omega. \tag{12}$$

The vector $\omega \in W$ is called the instantaneous angular velocity given by Eq. (13).

$$\omega = \psi e_3 + \theta E_2^{-2} + \phi E_1^{-1}. \tag{13}$$

In numerous studies and texts, the angular velocity vector $(\psi, \theta, \phi)$ of the Tait-Bryan angles [4], [5] is often referred to as coordinate components $(\dot{\psi}, \dot{\theta}, \dot{\phi})$ of the Tait-Bryan angles [4], [5]. Hence, the base vectors of the moving coordinate system, and $\dot{I}$ be the moment of inertia of the quadcopter. Using operator $B$, we obtained the following equations:

$$h = \dot{I} \omega = BH \in w, \tag{15}$$

$$H = \dot{I} \Omega \in W. \tag{16}$$

In addition, let $r \in w$ be the torque of the quadcopter. We obtain the time derivative of an angular momentum which is equal to the moment, as follows:

$$\frac{d}{dt}h = \tau = \frac{d}{dt}BH + \dot{B}H + \dot{\Omega} [\Omega, H] - T = 0. \tag{17}$$

Then, $\Omega$ and $\dot{\Omega}$ can be expressed in concrete terms by:

$$\Omega = (-\dot{\psi} \sin \theta + \dot{\phi}) e_1 + (\psi \cos \theta \sin \phi + \dot{\theta} \cos \phi) e_2 + (\psi \cos \theta \cos \phi - \dot{\phi} \sin \phi) e_3, \tag{18}$$

and

$$\dot{\Omega} = (-\dot{\phi} \sin \theta - \dot{\psi} \cos \phi + \dot{\phi}) e_1 + (\dot{\psi} \cos \theta \sin \phi + \dot{\phi} \cos \phi + \dot{\psi} \sin \phi) e_2 + (\dot{\phi} \sin \theta \cos \phi - \dot{\phi} \sin \phi) e_3. \tag{19}$$

respectively.

A torque $\tau$ of Eq. (17) is also given as:

$$\tau = [r, f]. \tag{20}$$
Here, \( \mathbf{r} \) and \( \mathbf{f} \) denote the position vector on which external forces act and a vector of the external forces, respectively. The coordinate components of an angular momentum of the quadcopter are given as Eq. (21).

\[
\mathbf{H}_A = \sum_{i=1}^{3} H_{Ai} \mathbf{E}_i. \quad (21)
\]

Then, we have the following equation:

\[
(H_1, H_2, H_3)^T = \hat{\mathbf{I}} (\Omega_1, \Omega_2, \Omega_3)^T. \quad (22)
\]

The moment of inertia \( \hat{\mathbf{I}} \) is also defined by

\[
\hat{\mathbf{I}} = \begin{bmatrix}
I_{11} & I_{12} & I_{13} \\
I_{12} & I_{22} & I_{23} \\
I_{13} & I_{23} & I_{33}
\end{bmatrix}. \quad (23)
\]

In this paper, for \( \hat{\mathbf{I}} \), we assume that \( I_{ij} = 0 \), and for \( k \neq l \).

3. Dynamic Model of a Quadcopter

The coordinate systems and free body diagram for the quadcopter are shown in Figure 3. Based on the preceding mathematical foundation, we can describe the dynamic model of the quadcopter (Fig 3).

![Fig 3: Coordinate systems and forces/moments acting on the quadcopter (\( m = 1.656 \text{[kg]} \), \( g = 9.80665 \text{[m/s}^2] \), \( I_{11} = 0.01982 \text{[kg} \cdot \text{m}^2] \), \( I_{22} = 0.01954 \text{[kg} \cdot \text{m}^2] \), \( I_{33} = 0.03221 \text{[kg} \cdot \text{m}^2] \), \( L = 0.365 \text{[m]} \)]](image)

\( F_i, (i = 1, 2, 3, 4) \) and \( M_i, (i = 1, 2, 3, 4) \) of Figure 3, represent vertical forces and moment, respectively. \( F_i, (i = 1, 2, 3, 4) \) and \( M_i, (i = 1, 2, 3, 4) \) are defined in a similar manner [1]. Each motor of the quadcopter has an angular speed \( \omega_i \) and produces a vertical force \( F_i \) according to:

\[
F_i = k_F \omega_i^2, \quad i = 1, 2, 3, 4, \quad (24)
\]

Experimentation with a fixed motor in a steady state shows that \( k_F \approx 1.79 \times 10^{-3} \text{[N m/rev]} \). The motors also produce a moment according to:

\[
M_i = k_M \omega_i^2, \quad i = 1, 2, 3, 4, \quad (25)
\]

The constant, \( k_M \), is determined to be approximately \( 4.38 \approx 10^{-3} \text{[N m/rev]} \) by matching the performance of the simulation to the real system.

We establish the Lagrangian of the quadcopter \( \text{Lag} \), as follows:

\[
\text{Lag} = \frac{1}{2} m(\dot{\mathbf{r}}, \dot{\mathbf{r}}) + \frac{1}{2} \left( \hat{\mathbf{I}} \dot{\mathbf{\Omega}}, \mathbf{\Omega} \right) - m g \mathbf{r}_3, \quad (26)
\]

where \( \left( \cdot, \cdot \right) \): the scalar product, \( \mathbf{r} = (r_1, r_2, r_3) \).

We also identify Lagrange’s equations for the quadcopter \((k = 1, 2, 3), \) as follows:

\[
\frac{d}{dt} \frac{\partial \text{Lag}}{\partial \dot{\mathbf{r}}_k} - \frac{\partial \text{Lag}}{\partial r_k} = \left( \mathbf{B} \sum_{i=1}^{4} \mathbf{F}_i, \mathbf{e}_k \right). \quad (28)
\]

We summarize the vector equations of the quadcopter, as follows:

\[
\hat{\mathbf{I}} \dot{\mathbf{\Omega}} + [\mathbf{\Omega}, \hat{\mathbf{I}} \dot{\mathbf{\Omega}}] = L(F_4 - F_1) \mathbf{E}_1 + L(F_2 - F_3) \mathbf{E}_2 \\
+ (M_1 - M_2 - M_3 + M_4) \mathbf{E}_3, \quad (29)
\]

\[
m \ddot{\mathbf{r}} = -mg \mathbf{r}_3 + \mathbf{B} \sum_{i=1}^{4} \mathbf{F}_i, \quad (30)
\]

\[
\hat{\mathbf{I}} \mathbf{E}_k = I_{1k} \mathbf{E}_1 + I_{2k} \mathbf{E}_2 + I_{3k} \mathbf{E}_3, \quad (31)
\]

where \( m \): the real weight of the quadcopter.

Since the operator \( \mathbf{B} \) states variables \( \mathbf{\Omega} \) and \( \dot{\mathbf{\Omega}} \) are expressed as the functions of \((\psi, \theta, \phi, \psi, \theta, \phi)\), the state equation of the quadcopter can be rewritten as equations of the function of \((\psi, \theta, \phi, \psi, \theta, \phi)\).

\[
\mathbf{x} = \mathbf{F}(\mathbf{x}) + \mathbf{G}(\mathbf{x}, \mathbf{u}), \quad (32)
\]

\[
\mathbf{x}^T = (\psi, \theta, \phi, \psi, \theta, \phi), \quad (33)
\]

\[
\delta \mathbf{x}^T = (\delta \psi, \delta \theta, \delta \phi, \delta \psi, \delta \theta, \delta \phi), \quad (34)
\]

\[
\mathbf{u}^T = (\omega_1^2, \omega_2^2, \omega_3^2, \omega_4^2), \quad (35)
\]

\[
\mathbf{F}(\mathbf{x}) = \begin{bmatrix}
f_\psi(\psi, \theta, \phi, \theta, \phi) \\
f_\theta(\psi, \theta, \phi, \theta, \phi) \\
f_\phi(\psi, \theta, \phi, \theta, \phi)
\end{bmatrix}, \quad (36)
\]
The variational equation (38) that can be used to control the quadcopter is described by

$$\frac{d}{dt} \delta x = DF(x_0) \cdot \delta x + B_G u,$$

(38)

where $DF(x_0)$ and $x_0$ denote the Jacobian of $F(x)$ and a driving point of $x$. In particular, $f_\theta$, $f_\phi$, and $f_\psi$ in $F(x)$ are easily obtained using Maple symbolic computations on $B_{G \theta}^T$, $B_{G \phi}^T$, and $B_{G \psi}^T$, as follows:

$$B_{G \theta}^T = [-I_{22} \cos \phi \cos \theta \sin \theta \phi_k + I_{22} \cos \phi \cos \theta \sin \theta \phi_k + I_{22} \cos \phi \cos \theta \sin \theta \phi_k + I_{22} \cos \phi \cos \theta \sin \theta \phi_k] \sin \kappa_1 + I_{22} \cos \phi \cos \theta \sin \theta \phi_k - I_{22} \cos \phi \cos \theta \sin \theta \phi_k - I_{22} \cos \phi \cos \theta \sin \theta \phi_k] \sin \kappa_1,$$

$$B_{G \phi}^T = [-I_{22} \sin \phi \kappa_M + I_{22} \sin \phi \kappa_M + I_{22} \sin \phi \kappa_M + I_{22} \sin \phi \kappa_M] \sin \kappa_1 + I_{22} \cos \phi \cos \theta \sin \theta \phi_k - I_{22} \cos \phi \cos \theta \sin \theta \phi_k - I_{22} \cos \phi \cos \theta \sin \theta \phi_k - I_{22} \cos \phi \cos \theta \sin \theta \phi_k] \sin \kappa_1,$$

$$B_{G \psi}^T = [-I_{22} I_{33} \sin \phi \kappa_4 - I_{22} I_{33} \sin \phi \kappa_4 - I_{22} I_{33} \sin \phi \kappa_4 - I_{22} I_{33} \sin \phi \kappa_4],$$

(39) (40) (41)

4. Simulation of Motion of a Quadcopter

By means of numerical computations on MATLAB with ode45 solver $t \in [0 \ 3]$ applied to Eq. (30) and Eq. (32), an example of typical quadcopter flight behavior without any controls is obtained. This behavior includes 3m hovering, forward moving and descending, and crashing. The simulation results illustrated in Figure 4 and Figure 5 are obtained. The initial values are set as $\psi = \theta = \phi = 0 \ [\text{rad/s}]$, $\psi = \theta = \phi = 0 \ [\text{rad}]$, $r_1 = r_2 = 0 \ [\text{m/s}]$, $r_1 = r_2 = 0 \ [\text{m/s}]$, $r_1 = 3 \ [\text{m}]$, $\omega_M = \omega_M = \omega_M = 4760 \ [\text{rpm}]$, and $\omega_M = 4770 \ [\text{rpm}]$. Notice that the numerical computation of Eq. (30) is carried out by using the computation results of Eq. (32), together with interpolations of $\psi(t)$, $\theta(t)$ and $\phi(t)$.

5. Conclusion

The following results are obtained:

(1) We have derived the system dynamics and system control equations for the quadcopter by using the operator $B$ together with two coordinate systems, $w$ and $W$. In addition, we have described Lagrange’s equations for the quadcopter.

(2) We have reliably illustrated the typical behavior of the quadcopter using numerical simulations on MATLAB with an ode45 solver.

References


Robust performance of two-wheeled mobile robot circular formations controlled by coupled oscillators

Tadashi Nakamura†, Koki Yoshida‡, Keiji Konishi‡, and Naoyuki Hara†

†Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531 JAPAN
Email: konishi@eis.osakafu-u.ac.jp

Abstract—Formation control has recently received considerable attention in the field of robotics and control theory. The present paper deals with two-wheeled mobile robots which achieve circular formations based on dynamics of coupled oscillators. The main advantage of this control is that circular formations can be obtained by a simple control law based on nonlinear dynamics. The purpose of this paper is to show another advantage, robustness of formation: even if some robots on formations stop due to trouble, the remaining robots make up for the troubled robots and keep to form the circular pattern without changing control law. It is shown on numerical simulations that the robustness of formations depends on how we choose robots to be removed from formed robots and to be added to them.

1. Introduction

Considerable attention has been paid to the research of complex network science in a variety of fields. Many subjects on this research are based on dynamics of coupled oscillators [1]. In the past decades, the collective phenomena in coupled oscillators have been received more and more attention because they provide us useful information on mechanism of various nonlinear behaviors in complex networks [2, 3]. Nowadays, their potential applications in engineering field have been demonstrated in many fields: Fukunaga et al. proposed a new control system for reducing peak power in energy storages on the basis of collective behavior in oscillators coupled by delayed power price [4]; Okuda et al. used synchronization of pulse-coupled oscillators to synchronize wireless sensor networks [5]; Zhou and Low employed collective behavior of coupled oscillators for locomotion control of an underwater vehicle [6].

In recent years, formation control, which manages behavior of multiple robots such that they form a specific pattern, has been one of the hot topics in the field of robotics and control theory. Hara et al. showed that a simple control law based on dynamics of coupled oscillators achieves mobile robot circular formations [7]. Tsukiji et al. experimentally demonstrated that the control law works well for two-wheeled mobile robots [8]. Very recently, Nakamura et al. [9] provided a simple approach for analyzing stability of mobile robot circular formations controlled by the simple control law [7,8]. This approach gave us a simple design procedure of control parameters stabilizing a desired formation.

Indeed, from a practical viewpoint, we should consider a natural situation where some robots stop due to trouble. Even for such situation, it is expected that the remaining robots make up for the troubled robots and keep to form the circular pattern. Although the previous studies [7–9] dealt with two-wheeled mobile robot circular formations for a given number of robots, it remains an unsettled question how robots cope with such situation. The purpose of the present paper is to investigate behavior of two-wheeled mobile robots in a situation where some robots are removed from and added to circular formations during their work. It is shown that the circular formations are robust against small remove/add disturbances and large balanced remove/add disturbances. However, we notice that large unbalanced remove/add disturbances may destroy the circular formations, and induce other pattern formations.

2. Two-wheeled mobile robots [9]

As illustrated in Fig. 1, dynamics of two-wheeled mobile robot $i \in \{1, \ldots, N\}$ can be expressed as

$$\begin{bmatrix}
\dot{r}_i \\
\dot{\theta}_i
\end{bmatrix} = \begin{bmatrix}
\cos \theta_i & 0 \\
\sin \theta_i & 0
\end{bmatrix} \begin{bmatrix}
v_i \\
\omega_i
\end{bmatrix}.$$  \hspace{1cm} (1)

The robot $i$ moves with the angular velocity of robot $i$ around the origin, $\kappa_i \in \mathbb{R}$, and is located distance $r_i > 0$ from the origin. The robot $i$ moves with the angle $\theta_i \in \mathbb{R}$ to the radial direction. $\psi_i \in \mathbb{R}$ defines the angle between the radial direction of robot $i$ and that of $i+1$. Here the robot $i$ is forced by the control signals, the heading-direction component of velocity $v_i \in \mathbb{R}$ and the angular velocity around its center $\omega_i \in \mathbb{R}$. We obtained a desired control signals by adjusting the rotational velocity of its two wheels.

Let us move on to reference dynamics of one-way coupled oscillators,

$$\dot{r}_i = f(r_i, \dot{r}_i) := ar_i \left(1 - \frac{r_i^2}{\bar{r}_i^2}\right),$$  \hspace{1cm} (2a)

$$\kappa_i = g(\psi_i) := \Omega + \kappa \sin \psi_i.$$  \hspace{1cm} (2b)

Equations (2a) and (2b) describe dynamics of $i$-th distance $r_i$ and dynamics of one-way coupled phase oscillators. As
the angle between the radial directions for robots \( N \) and \( 1 \) is denoted by \( \psi_i \), one-way coupled phase oscillators (2b) has the periodic boundary. The coupled oscillators (2) have the following parameters: \( \bar{r}_i > 0, a \in \mathbb{R}, \Omega \geq 0, \) and \( \varepsilon \in \mathbb{R} \).

Our previous study [9] proposed control law for \( v_i \) and \( \omega_i \), which leads the two-wheeled mobile robots (1) to behave of the reference dynamics,

\[
\begin{align*}
    \dot{v}_i &= \bar{v}_i := k_v \{ f (r_i, \hat{r}_i) \cos \theta_i + r_i g (\psi_i) \sin \theta_i \}, \quad (3a) \\
    \dot{\omega}_i &= \bar{\omega}_i := k_\omega \{ r_i g (\psi_i) \cos \theta_i - f (r_i, \hat{r}_i) \sin \theta_i \}. \quad (3b)
\end{align*}
\]

The feedback gains, \( k_v \in \mathbb{R} \) and \( k_\omega \in \mathbb{R} \), can be set as one wants.

The control law (3) suggests that the robot \( i \) always measures the three real-time data, \( r_i, \theta_i, \) and \( \psi_i \). Since every robot is supposed to be able to measure the distance to a target and the angle between targets in real time, they can autonomously move in accordance with control law (3).

The state space model of \( N \) robots (1) controlled by law (3) with \( \psi_i := \kappa_{i+1} - \kappa_i \) is given by

\[
\begin{align*}
    \dot{r}_i &= \bar{r}_i \cos \theta_i \\
    \dot{\theta}_i &= \bar{\omega}_i, \quad (i = 1, \ldots, N) \\
    \dot{\psi}_i &= \frac{1}{r_{i+1}} \bar{r}_{i+1} \sin \theta_{i+1} - \frac{1}{r_i} \bar{r}_i \sin \theta_i, \quad (i = 1, \ldots, N - 1).
\end{align*}
\]

This model has equilibrium points,

\[
r_{i} = \hat{r}_{i}, \quad \theta_{i} = \frac{\pi}{2}, \quad \psi_{i} = 2\pi \frac{l}{N},
\]

where \( l \in \{0, \ldots, N\} \) denotes a type of formations. The equilibrium point with formation \( l \) indicates that robots \( i \in \{1, \ldots, N\} \) move at equally spaced intervals on circles with radius of \( \hat{r}_{i} \). Our previous study provided a procedure for design of control parameters.

\[
\begin{array}{|c|c|c|c|}
\hline
\kappa_\omega & a > 0 & \varepsilon > 0 & \varepsilon < 0 \\
\hline
k_v > 0 & k_v > 0 & k_v > 0 & k_v < 0 \\
\hline
\end{array}
\]

Fact 1 ([9]). Formation number \( l \), total number of robots \( N \), and angular velocity \( \Omega \geq 0 \), are assumed to be given. If the parameters in control law (3) \( \{ \varepsilon, a, k_v, k_\omega \} \) are designed in accordance with Table 1, then formation \( l \) \( \{ \text{i.e., equilibrium point (5) with formation } l \} \) is stable.

It should be noted that, for the designed parameters, the other formations might be stabilized because Table 1 is derived on the basis of sufficient condition for equilibrium point (5) with formation \( l \) to be stable.

3. Removing/Adding of robots

This section investigates behavior of circular formation \( l = 1 \) when some robots are removed or added. From Fact 1, we can obtain the following result.

Corollary 1. If the parameters in control law (3) are set to

\[
k_\omega > 0, \varepsilon > 0, k_v > 0, a > 0,
\]

then formation \( l = 1 \) \{ i.e., equilibrium point (5) with formation \( l = 1 \) \} with counterclockwise direction is stable for any \( N \geq 5 \).

Proof. It is clear from Table 1 that formation \( l = 1 \) on \( N \) robots with designed parameters (6) is stable for

\[
\frac{1}{N} \in \left[ 0, \frac{1}{4} \right] \iff 4 < N.
\]

This fact indicates that if the number of robots is equal to or greater than five, formation \( l = 1 \) is stable independent of the number. In addition, the angular velocity at formation \( l = 1 \) with the designed parameters (6) is given by

\[
k_v = \frac{\bar{v}_i}{\bar{r}_i} \sin \frac{\pi}{2} = k_v \left\{ \Omega + \varepsilon \sin \left( 2\pi \frac{l}{N} \right) \right\} > 0,
\]

for all \( i \in \{1, \ldots, N\} \). The positive \( k_v \) indicates that robot \( i \) runs with counterclockwise direction.

This corollary implies that formation \( l = 1 \) remains stable even if the number of robots changes with time under condition \( N \geq 5 \). In other words, the robots can remain to form \( l = 1 \) if some robots are removed or added under this condition.
4. Numerical examples

This section will provide some numerical examples to verify the analytical results. In our numerical examples, small uniformly distributed random signals with amplitude $[-1.0 \times 10^{-4}, 1.0 \times 10^{-4}]$ are added to the right-hand side of control signals (3a) (3b) in order to confirm the local stability of formations.

We consider the following situation: $N = 10$ robots run with formation $l = 1$; then, two of them are removed (i.e., $N : 10 \rightarrow 8$); finally, two robots are added (i.e., $N : 8 \rightarrow 10$). Corollary 1 guarantees that formation $l = 1$ remains stable for such removing/adding of robots if the parameters in control law (3) are set in accordance with Eq. (6). Figure 2 shows time series data of angles $\psi_i$ ($i = 1, \ldots, 10$). It can be seen that the robots 9 and 10 are removed at $t = 100$, and then the eight remaining robots keep formation $l = 1$ with $\psi_i = \pi/4$ ($i = 1, \ldots, 8$) through transient behavior. The two robots are added at $t = 200$, and then the ten robots keep formation $l = 1$ with $\psi_i = \pi/5$ ($i = 1, \ldots, 10$) through transient behavior.

Corollary 1 states that, for $N = 10$, formation $l = 1$ remains stable if we remove and add up to five robots. Now we remove and add three robots. The time series data of angles $\psi_i$ ($i = 1, \ldots, 10$) are shown in Fig. 4. The robots 8, 9, and 10 are removed at $t = 100$. However, the seven remaining robots change their formation from $l : 1 \rightarrow 0$ through transient behavior. Remark that Fact 1 suggests the coexistence of formations $l = 0$ and $l = 1$. Further, the three robots are added at $t = 200$, but the ten robots cannot turn back to $l = 1$.

In order to avoid destroying formation $l = 1$, we choose the three robots (3, 6, and 9) uniformly from ten robots.
Figure 4: Time series data of angles $\psi_i$ ($i = 1, \ldots, 10$) in the case that three robots (i.e., robots 3, 6, and 9) are removed at $t = 100$ and are added at $t = 200$.

Acknowledgments

This research was partially supported by JSPS KAKENHI (26289131).

References

Modeling and Stabilization of the Novel Quadrotor with Tilting Propeller

Takumi Fukuda,† Akinori Sakaguchi,† Takashi Takimoto,‡ and Toshimitu Ushio†

†Graduate School of Engineering Science, Osaka University
1–3 Machikaneyama, Toyonaka, Osaka, 560-8531 Japan
‡Department of Creative Engineering, National Institute of Technology, Kitakyushu College
5–20–1 Shii, Kokuraminamiku, Kitakyushu, Fukuoka, 802-0985 Japan
Email: fukuda, sakaguchi@hopf.sys.es.osaka-u.ac.jp, takashi@kct.ac.jp, ushio@sys.es.osaka-u.ac.jp

Abstract—The quadrotors are useful for periodic inspection of tunnels and bridges. However, the conventional quadrotor has a problem such that it is impossible to control the rotational motion and the translational motion independently. Therefore, if the camera is attached on the bottom of the quadrotor, it is difficult to observe its upward direction. In this paper, we develop a novel quadrotor that has a link to tilt its propellers. By increasing the degree of freedom of the quadrotor, we resolve the problems described above. And we derive a model of the developed quadrotor. Then, we consider a PID controller for the stabilization of the quadrotor at a specified hovering state. We derive gain parameters of the PID controller by which the controlled quadrotor is stabilized.

1. Introduction

The development of a quadrotor has been drastically evolved over the last decade. Since its structure is simple and it has high mobility, maintainability, and inexpensive-ness, it has been utilized in many fields such as the surveillance and the exploration of disasters (such as a fire, an earthquake, and a flood), and periodic inspection of bridges and tunnels[1, 2].

We consider the case of the inspection of a bridge or a tunnel. Then, the quadrotor observes both its lateral and upward direction. Usually, the camera is attached on the bottom of the quadrotor. So, it is easy to observe its lateral direction, but it is difficult to observe its upward direction because the camera cannot turn to upward. If we attach the camera on the upward of the quadrotor, it is difficult to observe the downward. Moreover, we cannot control the attitude and translational motion of the quadrotor independently because it is an underactuated mechanical system. In other words, a quadrotor has 6 degrees of freedom (3 dimensional translational motions and 3 dimensional attitude rotations) with only 4 degrees of freedom control inputs (the thrust, roll input, pitch input, and yaw input).

Many approaches to the increase of the degrees of its freedom have been studied [3, 4, 5]. M. Ryll et al. developed a quadrotor that has 6 degrees of freedom. They realized that the mounting frame of the rotor that is able to tilt, and the quadrotor has 8 inputs. [6, 7]. However, it is impossible to tilt the quadrotor largely enough to observe the upward. A. Oosedo et al. developed a quadrotor that can change the pitch angle at hovering from 0 to 90 degrees by further increasing the angle of tilt [3]. P. Segui-Gasco et al. developed a quadrotor that has the 8 inputs. By doing so, it does not fall immediately even if one of the rotor is broken [8]. K. Kawasaki et al. developed an H-shaped quadrotor. It can carry out the rotation of the pitch angle of 360 degrees. It needs two additional servo motors for tilting [9].

In this paper, we develop a novel quadrotor that named a parallel linked quadrotor shown in Fig. 1. We utilize a parallel link at the frame of the x-axis direction, and install one servo motor at the center of the quadrotor so that its pitch angle can move from -90 degree to 90 degree. We call this angle a tilt angle. An advantage of this novel quadrotor is that it can hover on the spot in any tilt angle just by using one servo motor. On the other hand, the previous studies used two or more servo motors. Thus, the developed quadrotor has ease of maintenance. In Section 2, we derive the model of developed quadrotor. In Section 3, we consider a PID controller for the stabilization of the developed quadrotor at a specified hovering state. And we investigate gain parameters of the PID controller by which the controlled quadrotor is stabilized. Finally, Section 4 concludes the paper.
the body. As shown in Fig. 3, each propeller from the front in a clockwise is named the numbers 1 to 4. The propellers 1 and 3 rotate the counter-clockwise while the propellers 2 and 4 rotate the clockwise. They produce thrust power by rotating all propellers. The rotational movements of the quadrotor are called roll, pitch, and yaw respectively. Roll is rotational motion around the x’ axis of the quadrotor. In the same way, pitch and yaw are rotational motions around the y’ and the z’ axis of the quadrotor, respectively. In addition to these general rotational movements, we introduce a tilt angle $\alpha$ in the developed quadrotor as shown in Fig. 2. This angle moves by only one servo motor. On the other hand, the thrust and the anti-torque of each rotor are known to be proportional to the square of the rotational speed of the motor. Since the direction of the rotation of the propellers 1 and 3 is opposite to that of the propellers 2 and 4, the corresponding anti-torques generated by the propellers are also in the opposite directions. The thrust $T_i$ and the anti-torque of each rotor $Q_i$ are described by the following equations.

$$T_i = b\Omega_i^2,$$

(1)

$$Q_i = (-1)^ib\Omega_i^2,$$

(2)

where $i = 1, 2, 3, 4$ are propellers’ numbers, $b$ is a thrust constant, and $d$ is an anti-torque constant. Each propeller is driven by a motor, and the thrusts $T_i$ are generated by the propeller rotations. In the quadrotor model, we consider the total thrust $U_1[N]$ and the torque around each axis of the body coordinate system $U_2, U_3,$ and $U_4[Nm]$ as inputs, which are described by the following equations. They are roll, pitch, and yaw input, respectively.

$$U_1 = b(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2),$$

(3)

$$U_2 = lb(-\Omega_1^2 + \Omega_3^2),$$

(4)

$$U_3 = (l_1 \cos \alpha + l_2)b(-\Omega_1^2 + \Omega_3^2),$$

(5)

$$U_4 = d(-\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2),$$

(6)

where $l_1$ and $l_2$ are the lengths of the arms as shown in Fig. 2. From (5), the tilt angle $\alpha$ has an effect on the pitch input $U_3$ only.

$$\Gamma_E = [x \ y \ z]^T$$

and $\Theta_E = [\phi \ \theta \ \psi]^T$ denote a position vector of the center of gravity and an altitude angle vector in the E-frame, respectively. $\phi$, $\theta$, and $\psi$ are called a roll, a pitch, and a yaw angle, respectively. $V_B = [v_x \ v_y \ v_z]^T$ and $\omega_B = [\omega_x \ \omega_y \ \omega_z]^T$ denote a velocity vector in the B-frame, respectively. The kinematics of the developed quadrotor is described as follows.

$$\Gamma_E = RV_B,$$

(7)

$$\Theta_E = T\omega_B,$$

(8)

where $R$ is a rotation matrix and $T$ is a transfer matrix given by

$$R = \begin{bmatrix}
c_{\psi}c_{\theta} & c_{\psi}c_{\phi} + s_{\psi}s_{\phi}c_{\theta} & s_{\psi}c_{\phi} + c_{\psi}s_{\phi}c_{\theta} \\
c_{\psi}c_{\phi} + s_{\psi}s_{\phi}c_{\theta} & -s_{\psi}c_{\theta} + c_{\psi}s_{\phi}c_{\phi} & -c_{\psi}c_{\phi} \\
-s_{\psi} & c_{\phi}s_{\theta} & c_{\phi}c_{\theta}
\end{bmatrix},$$

(9)
Table 1: Specifications of the parallel linked quadrotor.

<table>
<thead>
<tr>
<th>Value</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m )</td>
<td>1.46 [kg]</td>
</tr>
<tr>
<td>( l_1 )</td>
<td>0.220 [m]</td>
</tr>
<tr>
<td>( l_2 )</td>
<td>0.165 [m]</td>
</tr>
<tr>
<td>( l )</td>
<td>( l_1 + l_2 ) [m]</td>
</tr>
<tr>
<td>( J_m )</td>
<td>( 2.06 \times 10^{-4} ) [kg m^2]</td>
</tr>
<tr>
<td>( J_{x0} )</td>
<td>( 2.41 \times 10^{-3} ) [kg m^2]</td>
</tr>
<tr>
<td>( J_{y0} )</td>
<td>( 4.51 \times 10^{-3} ) [kg m^2]</td>
</tr>
<tr>
<td>( J_{z0} )</td>
<td>( 2.51 \times 10^{-3} ) [kg m^2]</td>
</tr>
</tbody>
</table>

\[ T = \begin{bmatrix} 1 & s_{\phi} t_{\theta} & c_{\phi} t_{\theta} \\ 0 & c_{\phi} & -s_{\phi} \\ 0 & s_{\phi}/c_{\theta} & c_{\phi}/c_{\theta} \end{bmatrix}, \quad (10) \]

and \( s_j = \sin j \), \( c_j = \cos j \), \( t_j = \tan j \) \((j = \phi, \theta, \psi)\). Thus, the quadrotor is modeled by the following equation.

\[
\begin{bmatrix}
  m J_{3 \times 3} & 0_{3 \times 3} & 0_{3 \times 3} \\
 0_{3 \times 3} & J(\alpha) & V_B \\
0_{3 \times 3} & \omega_B \times (m V_B) & \omega_B \times (J(\alpha) \omega_B)
\end{bmatrix} = \begin{bmatrix} F_B \\
0 \quad 0 \quad 0 \end{bmatrix}, \quad (11)
\]

where \( I \) is the unit matrix, \( J(\alpha) = [J_1(\alpha) \ J_4(\alpha) \ J_6(\alpha)]^T \) is an inertia matrix of the developed quadrotor, and \( m \) is its mass. When the tilt angle \( \alpha \) changes, the shape of the developed quadrotor is also changed. Thus, the inertias depend on the tilt angle. The second term in the left-hand side is the Coriolis force acting on the apparent in B-frame. Besides, the translational motion is modeled by the E-frame. Its right-hand side represents forces and torques acting to the quadrotor. Concretely speaking, they represent gyro effects generated by the propeller rotations, the gravity force applied to the body, thrust forces and torques of propellers' rotations. In addition, the anti-torque is generated by tilting movement. This term appears in the pitch equation. Here, we can translate \( \omega_B \) to \( \Theta_B \) by using the transfer matrix \( T \). However, since we consider the situation of hovering, the attitude angle is close to zero. Thus a transfer matrix can be approximated to the unit matrix. In other words, we can replace \( \omega_B \) with \( \Theta_B \). Then, we can neglect the effect of air resistance generated by the translational motion. Under these assumptions, (11) is rewritten as follows.

\[
\begin{align*}
\dot{x} &= a_x \frac{v_1}{m}, \\
\dot{y} &= a_y \frac{v_1}{m}, \\
\dot{z} &= -g + a_z \frac{v_1}{m}, \\
\dot{\phi} &= \frac{J_4(\alpha) - J_6(\alpha)}{J_6(\alpha)} \dot{\theta} \psi - \frac{J_6(\alpha)}{J_6(\alpha)} \theta \Omega + \frac{v_1}{J_6(\alpha)}, \\
\dot{\theta} &= \frac{J_4(\alpha) - J_6(\alpha)}{J_4(\alpha)} \dot{\phi} \psi - \frac{J_4(\alpha)}{J_4(\alpha)} \phi \Omega + \frac{v_1}{J_4(\alpha)} - \ddot{\alpha}, \\
\dot{\psi} &= \frac{J_4(\alpha) - J_6(\alpha)}{J_4(\alpha)} \dot{\phi} \theta + \frac{v_1}{J_4(\alpha)}.
\end{align*}
\]

where \( a_x = (s_{\phi} t_{\theta} \rho + c_{\phi} \theta \varphi), a_y = (-c_{\phi} t_{\theta} \rho + s_{\phi} \theta \varphi), \)

\( a_z = (c_{\phi} \rho), J_m \) is a motor inertia, and \( \Omega = \Omega_1 + \Omega_2 + \Omega_3 + \Omega_4 \).

3. PID control

In this section, we consider PID control for the stabilization of the developed quadrotor at the hovering state. Fig. 4 shows a block diagram of the PID controller, where \( r(t) = [z_m \ 0 \ 0 \ \psi_m]^T, e(t) = r(t) - y(t), U(t) = [U_1 \ U_2 \ U_3 \ U_4]^T, \)

\( y(t) = [z \ \phi \ \theta \ \psi]^T, K_p = \text{diag}[K_{p\phi} K_{p\theta} K_{p\psi}], K_i = \text{diag}[K_{iz} K_{i\phi} K_{i\theta} K_{i\psi}], \)

\( K_d = \text{diag}[K_{dz} K_{d\phi} K_{d\theta} K_{d\psi}] \). Since we use the PID controller, the target state is an equilibrium point of the controlled quadrotor, that is, the equilibrium point is given by \( z = z_m, \psi = \psi_m, z = \phi = \phi = \theta = \theta = \psi = 0 \). We consider the linearized system of the controlled quadrotor around the equilibrium point. The parameters of the developed quadrotor are given in Table 1. And nominal gains of the controller are listed in Table 2.

Table 2: Nominal gains of the PID controller.

<table>
<thead>
<tr>
<th>Gain</th>
<th>Gain</th>
</tr>
</thead>
<tbody>
<tr>
<td>( K_{p\phi} )</td>
<td>100</td>
</tr>
<tr>
<td>( K_{p\theta} )</td>
<td>50</td>
</tr>
<tr>
<td>( K_{p\psi} )</td>
<td>50</td>
</tr>
<tr>
<td>( K_{p\phi} )</td>
<td>50</td>
</tr>
<tr>
<td>( K_{p\theta} )</td>
<td>10</td>
</tr>
<tr>
<td>( K_{i\phi} )</td>
<td>10</td>
</tr>
<tr>
<td>( K_{i\theta} )</td>
<td>10</td>
</tr>
<tr>
<td>( K_{i\psi} )</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig. 7 shows the behavior of the mathematical model when the initial value of roll angle is 0.2[rad] (= 11.5[deg]).

4. Conclusion

In this paper, we developed a novel quadrotor named the parallel linked quadrotor, where tilting is realized by a parallel link and one servo motor. We derived a model of the quadrotor. We consider PID control for the stabilization of its hovering state. We investigated stabilization regions on gain parameter planes. It was shown that, as the tilt an-
gle increases, we have to increase the gains for pitch while we can decrease those for yaw. Because, by the increase of the tilt angle, $J_x$ increases while $J_z$ decreases. The experiment by a prototype of the quadrotor is future work. It is also future work to investigate an effect of the anti-torque generated by the servo motor on the pitch rotation.

Acknowledgments This research was supported in part by JSPS KAKENHI Grant Number JP24656262.

References


A Charge-Based SiC Power MOSFET Model Considering On-State Resistance

Rui Zhou, Michihiro Shintani, Masayuki Hiromoto, and Takashi Sato

Graduate School of Informatics, Kyoto University
Yoshida-hon-machi, Sakyo, Kyoto 606-8501, Japan
Email: paper@easter.kuee.kyoto-u.ac.jp

1. Introduction

Silicon Carbide (SiC) is considered as one of the most promising materials for realizing power converters that operate with high power density and at a high frequency. Accurate circuit simulation is crucial to optimally design power converters. The simulation accuracy greatly depends on the accuracy of the SiC power device model.

Fitting-based equations have been widely used as the SiC power MOSFET model [1]. These models are recognized to efficiently simulate the electrical behavior of the SiC power MOSFET. However, in such models, prediction of the physical phenomena, such as process variation and reliability of SiC, is difficult because the mathematical model does not contain physically meaningful parameters that actually varies in real devices. Also, the characteristics outside the fitting region may not be sufficiently accurate.

Recently, inversion charge-based transistor models have been successfully applied for designing silicon-based integrated circuits. The charge-based models can accurately reproduce behavior of silicon devices computationally efficiently [2, 3]. However, charge-based model has not yet been applied to the models of SiC vertically double diffused MOSFETs (VDMOSFETs). Model equations have to be modified to represent I-V characteristics. In particular, parasitic resistances associated with the vertical current flow in the vertical double diffused structure have to be accurately modeled.

In this paper, we propose an accurate SiC power MOSFET model based on the channel charge. The proposed model takes into account the bias dependence of the parasitic resistances of the VDMOSFET. The capacitance characteristics are represented by the analytic expressions proposed in [4]. The I-V, C-V, and transient characteristics of the proposed model have been validated through experiments using a commercial SiC device.

2. Charge-Based Model

The charge-based model is a compact device model that is widely used for simulating silicon-based MOSFETs. On the basis of single equation, the charge-based model can accurately and consistently predict the circuit performance [2, 3]. No concatenation of separate equations is required to represent full operational range of the device. This is one of the advantages over the conventional models, such as [5]. Another advantage of the charge-based model is its efficiency. The drain current is accurately obtained through an analytic expression that represents inversion charge density at the source and drain electrodes.

We adopt EKV model [2] as the basis of our proposed model. The drain current \( I_{DS} \) is represented as [6]:

\[
I_{DS} = \beta \int_{V_S}^{V_D} \frac{-Q_i}{C_{OX}} dV. \tag{1}
\]

Here, \( V_D \) and \( V_S \) are drain and source voltages, respectively. \( \beta = \mu C_{OX} \frac{W}{L} \) is the transmission coefficient, being \( L \) the channel length, \( W \) the channel width, \( \mu \) the carrier mobility, and \( C_{OX} \) the oxide capacitance per unit area. Although Eq. (1) is an accurate expression of the physical current flow, it involves numerical integration of the inversion charge \( Q_i \). Its computational cost is in general too high for use as the compact device model. In the EKV model, the inversion charges at the drain and source ends of the channels are approximated to facilitate analytic calculation. With this approximation, the drain current is obtained with a small calculation time.

The EKV model considers the inversion charge \( Q_i \) in both strong and weak inversion regions. When the body effect is sufficiently small, by solving the Poisson’s equation at the surface of the channel, the inversion charge is written as

\[
Q_{i}(\text{strong inv.}) \approx -nC_{OX}(V_P - V_{ch}), \tag{2}
\]

\[
Q_{i}(\text{weak inv.}) \approx -K_wC_{OX}U_T \exp \left( \frac{V_P - V_{ch}}{U_T} \right). \tag{3}
\]

Here, \( V_P \) is the pinch-off voltage, \( V_{ch} \) is the electric potential in the channel, and \( U_T \) is the thermal voltage. In addition, \( n \equiv 1 + \gamma/(2 \sqrt{r_0 + V_{ch}}) \) and \( K_w \equiv (n - \gamma)/(2 \sqrt{r_0 + V_{ch}}) \) are used.
1) \exp((\Psi_0 - 2\Phi_F)/(U_T))). \gamma \text{ is the substrate bias factor, } \Psi_0 \text{ is the channel surface potential at } V_{ch} = 0, \Phi_F \text{ is the Fermi potential, and } V_{T0} \text{ is the threshold voltage at the zero bias.}

The pinch-off voltage \(V_P\) is defined as follows:

\[
V_P = V_G - V_{T0} - \gamma \left[ \sqrt{V_G - V_{T0} + \frac{\gamma}{2}} + \frac{\gamma}{2} \right].
\]

(3)

Figure 1 illustrates the surface potential as a function of \(V_{ch}\) at a constant gate voltage \(V_G\). \(V_{ch}\) varies from zero to \(V_P\), and the surface potential also changes according to \(V_{ch}\).

The current that flows from the source to the drain (forward current \(I_F\)) is obtained by integrating \(Q_i/C_{OX}\) from \(V_S\) to infinity. Similarly, by integrating from \(V_S\) to infinity, the current from the drain to the source (reverse current \(I_R\)) is obtained. Because the drain current \(I_{DS}\) is the difference between \(I_F\) and \(I_R\), \(I_{DS}\) can be expressed as follows:

\[
I_{DS} = I_F - I_R = \beta \int_{V_S}^{\infty} \frac{-Q_i}{C_{OX}} dV - \beta \int_{V_S}^{\infty} \frac{-Q_i}{C_{OX}} dV.
\]

(4)

In order to calculate \(I_{DS}\) by Eq. (4), \(I_F\), \(I_R\), and the transition between \(I_F\) and \(I_R\) need to be modeled. \(I_F\) and \(I_R\) can be derived from Eq. (2). In the EKV model, the smoothing function is introduced to fit the transition.

3. Charge-Based SiC Power MOSFET Model

3.1. Overview of the Proposed Device Model

In this section, the proposed charge-based SiC power MOSFET model is explained. Figure 2 shows the structure of the proposed model. The drain current \(I_{DS}\) is expressed based on the charge-base model described in Sec. 2.

The proposed model assumes the VDMOSFET structure shown in Fig. 3. The VDMOSFET has the parasitic resistances on its current path, total of which is on-resistance \(R_{on}\). The value of each resistance is voltage dependent. Hence, all voltage dependencies of the on-resistances have to be correctly modeled in order to accurately simulate the current characteristic of the SiC power MOSFET.

The parasitic capacitances, \(C_{GS}, C_{DS}, C_{GD}\), are also modeled considering the dependences to the drain-source voltage \(V_{DS}\) and the gate-source voltage \(V_{GS}\). In the proposed model, the capacitance model proposed in [4] is used.

3.2. Drain Current Model

In order to apply the charge-based model to SiC VDMOSFETs, the following changes have been made.

- Mobility degradation due to interface traps at the SiC/SiO\(_2\) [7] has been considered.
- Bias-dependent on-resistances of VDMOSFET [8, 9] have been considered.

According to the measurement results of the drain current of SiC power MOSFETs, the transition between linear and saturation regions occurs more gradually than silicon transistors. The smooth transition is considered to be caused by high density of interface states in the inversion channel [7]. When body effect coefficient is sufficiently small, Eq. (3) can be rewritten as \(V_P = V_G - V_{T0}\). However, this equation is insufficient to express the gradual transition.

Through the observations on the measured I-V characteristics, we found that the gradual transition can be modeled by modifying the pinch-off voltage as:

\[
V_P = N_{GS}(V_{GS} - V_{T0})^\alpha + N_{DS}[(V_{GS} - V_{T0})^2V_{DS}]^\zeta,
\]

(5)

where \(N_{GS}, N_{DS}, \alpha, \text{ and } \zeta\) are fitting parameters.

Figure 3 shows the cross section of the VDMOSFET. The on-resistance \(R_{on}\) for this vertical structure is defined by considering the following resistances: source terminal resistance \(R_{source}\), drain terminal resistance \(R_{drain}\), substrate resistance \(R_{sub}\), accumulation resistance \(R_{acc}\), channel resistance \(R_{CH}\), JFET resistance \(R_{JFET}\), constant drift region resistance \(R_{drift-cons}\), and variable drift region resistance...
$R_{\text{drift-var}}$. Here, $R_{\text{source}}$, $R_{\text{drain}}$, $R_{\text{drift-cons}}$, and $R_{\text{sub}}$ are constant resistances determined by the geometry of the VDMOSFET.

$R_{\text{acc}}$ is the variable resistance. According to [8], due to the charge density of accumulation layer, $R_{\text{acc}}$ becomes $V_{\text{GS}}$ dependent

$$R_{\text{acc}} = \frac{1}{\mu C_{\text{ox}} (V_{\text{GS}} - V_{\text{TD}})}.$$  \hspace{1cm} (6)

$R_{\text{JFET}}$ is also the variable resistance formed at the JFET region. $R_{\text{JFET}}$ limits the drain current path by the width of the depletion layer that depends on $V_{\text{DS}}$. $R_{\text{JFET}}$ can be written as

$$R_{\text{JFET}} = \frac{\rho l}{L_{\text{D}} W}.$$  \hspace{1cm} (7)

$L_{\text{D}}$ is the distance between depletion layers, and it is defined as $L_{\text{D}} = L_{p} - 2x$. Here, $L_{p}$ is the distance between adjacent $p$ regions. $W$ is the channel width, $l$ is the vertical depth of the JFET region, and $\rho$ is the resistivity of the JFET region. $L_{\text{D}}W$ expresses the cross-sectional area of the JFET region. The width of the depletion layer $x$ is given as

$$x = \sqrt{\frac{2 \varepsilon_{\text{SiC}} N_{A} + N_{D}}{q} (V_{\text{bi}} + V_{\text{DS}})},$$  \hspace{1cm} (8)

where $V_{\text{bi}}$ is the built-in voltage, $\varepsilon_{\text{SiC}}$ is the permittivity of SiC, $q$ is the elementary charge, and $N_{D}$ and $N_{A}$ are the densities of donors and acceptors, respectively. By applying $V_{\text{DS}}$, $L_{\text{D}}$ changes. Combining the above equations, $R_{\text{JFET}}$ becomes:

$$R_{\text{JFET}} = \frac{\rho l}{L_{\text{D}} W} = \left(\frac{\rho l}{L_{p} - \sqrt{\frac{2 \varepsilon_{\text{SiC}} N_{A} + N_{D}}{q} (V_{\text{bi}} + V_{\text{DS}})}}\right) W.$$  \hspace{1cm} (9)

In the model of $R_{\text{drift-var}}$, $H_{1}$ and $H_{2}$ are defined as the depths from the substrate-surface of the device.

As shown in Fig. 3, by supposing $s$ as the cross-sectional area of the $R_{\text{drift-var}}$ between $H_{1}$ and $H_{2}$, $R_{\text{drift-var}}$ can be expressed as follows:

$$R_{\text{drift-var}} = \int_{H_{1}}^{H_{2}} \frac{\rho dh}{s} = W \left( L_{D} + \frac{L_{T} - 2L_{D}}{H_{2}} \right).$$  \hspace{1cm} (10)

where $L_{T}$ is the distance between adjacent source terminals. Therefore, Eq. (10) can be rewritten as follows:

$$R_{\text{drift-var}} = \int_{H_{1}}^{H_{2}} \frac{\rho dh}{W \left( L_{D} + \frac{L_{T} - 2L_{D}}{H_{2}} \right)} = C_{1} V_{\text{DS}} + \log(V_{\text{DS}}) + C_{2} V_{\text{DS}} + C_{3}.$$  \hspace{1cm} (11)

Here, $C_{1}$, $C_{2}$, and $C_{3}$ are the coefficients represented by using device dimensions and physical parameters.

The on-resistance $R_{\text{on}}$ of the VDMOSFET is considered as the series connection of all the above parasitic resistances. Thus, by arranging the sum of the parasitic resistances as $N_{1}$--$N_{s}$, $R_{\text{on}}$ is expressed as follows:

$$R_{\text{on}} = \frac{N_{1}}{V_{\text{GS}} - V_{\text{TD}}} + N_{2} V_{\text{DS}} + \frac{N_{3}}{N_{4} - \sqrt{V_{\text{DS}} + V_{\text{bi}}}} + N_{5}.$$  \hspace{1cm} (12)

### Table 1: Model parameters of current characteristic

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Explanation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W_{0}$ [V]</td>
<td>Surface potential when $V_{A} = 0$</td>
<td>6.22</td>
</tr>
<tr>
<td>$V_{DS}$ [V]</td>
<td>Threshold voltage at zero bias</td>
<td>5.31</td>
</tr>
<tr>
<td>$N_{GS}$</td>
<td>Coefficient of $V_{T}$ on $V_{GS}$</td>
<td>1.81×10^{-4}</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Multiplier of $V_{T}$ on $V_{GS}$</td>
<td>1.60</td>
</tr>
<tr>
<td>$N_{DS}$</td>
<td>Coefficient of $V_{T}$ on $V_{DS}$</td>
<td>7.52×10^{-4}</td>
</tr>
<tr>
<td>$\zeta$</td>
<td>Multiplier of $V_{T}$ on $V_{DS}$</td>
<td>1.13</td>
</tr>
<tr>
<td>$\mu C_{\text{ox}}$</td>
<td>Transconductance</td>
<td>3.61×10^{-3}</td>
</tr>
<tr>
<td>$N_{1}$</td>
<td>$\frac{q}{\varepsilon_{\text{SiC}} N_{A}}$</td>
<td>8.79×10^{-4}</td>
</tr>
<tr>
<td>$N_{2}$</td>
<td>Coefficient of $R_{\text{drift-var}}$</td>
<td>1.50×10^{-2}</td>
</tr>
<tr>
<td>$N_{3}$</td>
<td>$\rho l (\frac{\varepsilon_{\text{SiC}} N_{A}}{q})$</td>
<td>1.09</td>
</tr>
<tr>
<td>$N_{4}$</td>
<td>$C_{\text{ox}} (\frac{\varepsilon_{\text{SiC}} N_{A}}{q})$</td>
<td>41.27</td>
</tr>
<tr>
<td>$N_{5}$</td>
<td>Fixed resistance</td>
<td>2.09×10^{-1}</td>
</tr>
</tbody>
</table>

### Table 2: Model parameters of capacitance characteristic

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Explanation</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{bi}$ [V]</td>
<td>Built-in voltage</td>
<td>2.01</td>
</tr>
<tr>
<td>$V_{D}$ [V]</td>
<td>Threshold drain voltage</td>
<td>5.52×10^{-1}</td>
</tr>
<tr>
<td>$C_{GS}(0)$ [F]</td>
<td>$C_{GS}$ at zero bias</td>
<td>5.38×10^{-10}</td>
</tr>
<tr>
<td>$C_{DS}(0)$ [F]</td>
<td>$C_{DS}$ at zero bias</td>
<td>3.07×10^{-10}</td>
</tr>
<tr>
<td>$C_{GD}(0)$ [F]</td>
<td>$C_{GD}$ at zero bias</td>
<td>2.23×10^{-10}</td>
</tr>
<tr>
<td>$C_{\text{ox}}$ [F/Vs]</td>
<td>Gate-drain oxide capacitance</td>
<td>4.11×10^{-9}</td>
</tr>
</tbody>
</table>

### Figure 4: $V_{DS}$-$V_{GS}$ characteristics.

### 4. Experimental Results

The proposed device model is validated using a commercial SiC power MOSFET (1200V, 10A, [10]). I-V and C-V characteristics of the SiC power MOSFET are measured by a commercial curve tracer [11]. The proposed device model is implemented by Verilog-A. I-V, C-V, and transient characteristics are calculated by a commercial circuit simulator [12]. Model parameters are determined by a simulated annealing method [13]. The Model parameters used in the proposed model for I-V and C-V characteristics are summarized in Table 1 and 2, respectively.

#### 4.1. I-V and C-V Characteristics

The measured and simulated current and capacitance characteristics are compared in Figs. 4 and 5. In Fig. 4, $V_{GS}$ is varied from 6 V to 18 V with a 2 V step. In Fig. 5, capacitance values are measured from $V_{DS} = 0$ V to $V_{DS} =$
4.2. Transient Characteristic

The transient characteristic is evaluated using a double pulse tester circuit shown in Fig. 6. The switching frequency is 1 MHz and the pulse duty factor is 50%. The gate series resistance $R_g$ is varied as 33, 47, 68, and 100 Ω. Figure 7 shows both simulated (solid black) and measured (dashed red) waveforms at turn-on and turn-off periods. The proposed model can accurately simulate the switching waveforms of the SiC power MOSFET for all $R_g$ values.

But simulation wave pattern of $V_{DS}$, $I_{DS}$ is earlier than measurement. It is thought that this does not consider the parasitic ingredient in the actual survey circuit by the circuit simulation of this experiment.

5. Conclusion

In this paper, we proposed a charge-based model for SiC power devices considering the structure of the VDMOSFET. The proposed model also takes into account the gradual transition between linear and saturation regions due to interface traps. Experimental results using a commercial SiC power MOSFET show that the proposed simulation model accurately reproduces I-V, C-V, and transient characteristics.

Acknowledgments

This work has been partially supported by Kyoto Super Cluster Program (JST) and Cross-ministerial Strategic Innovation Promotion Program (SIP), “Next-generation power electronics” (NEDO).

References

A Design Example of Class-E Based Gate Driver for High Frequency Operation of SiC Power MOSFET
Michihiro Shintani†, Yuchong Sun‡, Hiroo Sekiya‡, and Takashi Sato‡

†Graduate School of Informatics, Kyoto University  ‡Graduate School of Advanced Integration Science, Chiba University

Abstract—Design of a class-E based gate driver has been presented for high frequency power converters using SiC power MOSFETs. Through numerical experiments, it is demonstrated that the proposed gate driver can eliminate its switching loss by zero voltage switching (ZVS) operation.

1. Introduction

Silicon Carbide (SiC) is a promising material for realizing high-frequency switching power converters [1]. Increasing the switching frequency allows us to reduce the size of converters. However, as switching frequency increases, switching loss of gate drivers becomes a big issue. Soft switching techniques, such as ZVS operation of class-E amplifiers [2], is one of the effective solutions to minimize the switching loss.

In this paper, we propose a novel gate driver based on class-E amplifier to drive SiC MOSFETs. By achieving the ZVS condition, the proposed gate driver eliminates the switching loss. Based on the equations described in [2], analytical equations for the gate driver application are also proposed to determine component values of the gate driver, such as conductors and inductors.

2. Gate driver design based on Class-E amplifier

Figure 1 shows the topology of the proposed gate driver for driving an SiC power MOSFET. It consists of dc supply voltage $V_{dd}$, dc-feed inductance $L_f$, shunt capacitance $C_s$, gate resistance $R_g$ and an ideal switch. $C_s$ is a capacitance to adjust capacitance in the resonance loop and $L$ is the inductance of the resonant circuit. As shown in Fig. 1, the SiC power MOSFET can be regarded as an equivalent circuit with an internal resistor $R_g$ and the input capacitance $C_{iss}$ (the sum of the drain-source capacitance $C_{ds}$ and the drain-gate capacitance $C_{gd}$). Thus equivalent capacitance becomes $C_s + C_{iss}$. The overall topology of the gate driver circuit is considered as a class-E amplifier. The switch turns on at zero voltage, and thus the energy loss of the gate driver can be reduced.

3. Design constraints

Component values of the proposed gate driver circuit can be determined basically by following the equations of class-E amplifiers. However, two constraints have to be additionally considered for driving the gate electrode of SiC power MOSFETs: (1) achieve resonance under the change of $C_{iss}$ depending on drain-source voltage $V_{ds}$ and gate-source voltage $V_{gs}$, (2) sufficiently large gate current $I_s$ to switch the SiC MOSFET.

By introducing $C_s$, these constraints can be simultaneously satisfied. With a small value of $C_s$, input capacitance can be made relatively constant. In the proposed design expressions, constant $\alpha < 1$ is introduced to scale $C_s$ as $C_s = \alpha \cdot C_{iss}$. Also, to turn on (off) the SiC MOSFET, $C_{iss}$ has to be fully charged (discharged). The required charge amount of $C_{iss}$ can be calculated once accurate $C_{iss}$ model is available, such as those in [3].

4. Numerical experiments

Figure 2 shows simulation waveforms of the example design driving a resistive load of 50$\Omega$ using SiC MOSFET model [3]. The drain voltage is 100$\mathrm{V}$, and the switching frequency is 10$\mathrm{MHz}$.

Component values are determined on the basis of the analytical equations. As shown in Fig. 2, there is no overlap between the switch voltage $V_s$ and the switch current $I_s$. The switching loss of the gate driver can be successfully eliminated by class-E configuration.

Acknowledgments This research is partially supported by Kyoto Super Cluster Program (JST) and Cross-ministerial Strategic Innovation Promotion Program (SIP), and “Next-generation power electronics” (NEDO).

Design Example of SiC Isolated Soft-Switching Driver

Ryoko Sugano†, Yuchong Sun†, Hiroo Sekiya†, and Takashi Hikihara‡

†Graduate School of Advanced Integration Science, Chiba University
1-33 Yayoi, Inage, Chiba 263-8522, Japan
‡Graduate School of Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615-8510, Japan
Email: r.sugano@chiba-u.jp, sekiya@faculty.chiba-u.jp

Abstract—This paper proposes an isolated resonant gate driver for SiC MOSFETs. By applying the class-E zero-voltage switching and zero-derivative switching (ZVS/ZDS) conditions, the proposed driver achieves high power-conversion efficiency at high frequencies. An analysis and a design example of the proposed driver are presented along with simulation and experimental results.

1. Introduction

Recently, Silicon Carbide (SiC) devices attract attentions as a next generation semiconductor devices [1]-[6]. SiC power devices can maintain high performance with high power. This is because SiC power devices have characteristics of high breakdown voltage and low thermal resistance. In most cases, SiC applications are considered at up to hundred kHz order. It is a challenging problem to use SiC devices at MHz-order frequencies.

The design of driver circuit is one of the important and difficult problems for operating SiC devices at high frequencies. Recently, gate drivers for high frequency operations of SiC MOSFETs [1]-[3] have been presented. The power loss at the driver is, however, cannot be ignored because of the hard switching. Additionally, the rectangular drive voltage strains due to parasitic capacitances and resistance on SiC devices.

It is one of the solutions to include the parasitic capacitance in the output resonant filter of the driver and the MOSFET is driven by the sinusoidal signal [7]-[10]. This is called as resonant driver. It is also an advantage of the resonant driver that the soft switching techniques can be adopted. Namely, the power-conversion efficiency is enhanced. It is, however, required that the output filter includes isolation topology for safety drive of the SiC.

This paper proposes an isolated resonant gate driver for SiC MOSFETs. The proposed driver includes a transformer for the isolations. In addition, the class-E zero-voltage switching and zero-derivative switching (ZVS/ZDS) conditions at turn-on instants are applied to the driver switch. Because of the class-E ZVS/ZDS conditions, the proposed driver achieves high power-conversion efficiency at high frequencies. An analysis of the proposed driver is carried out and a step-by-step design procedure based on the obtained analytical expressions is presented. As a design example, resonant SiC driver for 7 MHz operation is designed. The validity of the analytical expressions and design procedure were confirmed from the quantitative agreements with experimental and PSpice-simulation results.

2. Proposed Gate Driver

Figure 1 shows a proposed SiC gate driver. The basic topology is the class-E amplifier, which consist of input voltage \( V_p \), dc-feed inductance \( L_C \), MOSFET S as a switching device, shunt capacitance \( C_s \), and output filter. In the proposed driver, the isolation transformer is applied as shown in Fig. 1(a). The transformer is modeled the primary inductance \( L_1 \) and secondary inductance \( L_2 \) with mutual inductance \( M \) and equivalent series resistances (ESRs) of the coupling inductances \( r_{L1} \) and \( r_{L2} \). The primary side of the output filter is composed of \( C_1 \), \( C_p \) and \( L_1 \), where \( C_p \) is a component for adjusting the amplitude of the drive voltage. The secondary side of the output filter has \( L_2 \) and \( C_2 \), which is connected to gate of SiC MOSFET in series. The gate of the SiC is modeled as the gate capacitance and gate resistance, which are connected in series as shown in Fig. 1(b). Therefore, \( C_2 \) has a role to make an impedance matching at the secondary part. It is important to consider ESRs of the coupling inductances for considering the power-conversion efficiency optimization. Note that the gate voltage is the sum of the voltages across the \( C_s \) and \( R_g \).

Figure 2 shows example waveforms of the proposed gate driver, where \( \theta = \omega t = 2 \pi f t \) is the angular time and \( f \) is the operating frequency. When the switch is in ON state, the switch voltage is approximately zero. Conversely, current flows through the shunt capacitance when the switch is OFF state, which produces the pulse-type shape of switch voltage \( V_S \). The most important operation of the proposed driver is to achieve the class-E ZVS/ZDS conditions at the turn-on instant of switch voltage as shown in Fig. 2. Namely, the switch voltage satisfies

\[
v_S(2\pi) = 0 \quad \text{and} \quad \frac{dv_S(\theta)}{d\theta} \bigg|_{\theta=2\pi} = 0, \quad (1)
\]

simultaneously. Because of the class-E ZVS/ZDS conditions, the proposed driver achieves high power-conversion efficiency at high frequencies.

It is expected that a sinusoidal current flows through the output filter as shown in Fig. 2. This is because it is easy to control the duty ratio. For obtaining the sinusoidal output current, it is important to keep high loaded quality factor at the output filter. The gate capacitance \( C_s \) and gate resistance \( R_g \), however, are not small at SiC MOSFET, which is hard condition for achieving high \( Q \). This is a reason why the capacitance \( C_2 \) is added in the secondary part. By adding \( C_2 \), it is possible to reduce the resonant capacitance in the secondary part. Namely, high \( Q \) output
filter for obtaining a sinusoidal output current can be realized.

From above discussions, it is seen that the design of the secondary part is focused on the impedance matching. For gate-driver applications, it is important to adjust the output voltage of the gate signal for satisfying the maximum/minimum gate voltage conditions. Therefore, the impedance transform component $C_p$ is also mandatory component for designing the gate driver.

It is necessary to obtain a set of component values for achieving the class-E ZVS/ZDS conditions and specified amplitude of the gate voltage, which is a problem of the proposed driver designs. The circuit analysis is effective way to solve this problem.

3. Analysis of Proposed Driver

In this section, the analytical expressions of the proposed driver are given. The analytical expressions are necessary to design the proposed circuit with achieving the class-E ZVS/ZDS conditions.

3.1. Secondary Part

The current thorough the secondary part is expressed as

$$i_2 = I_2 \sin(\theta + \phi_r),$$

where $I_2$ and $\phi_r$ are the amplitude of $i_2$ and the phase shift between the driving signal of the switching device $D_r$ and the input current of the secondary part $i_2$. The amplitude of the secondary current is obtained from

$$I_2 = \frac{V_g}{jZ_{gj}} = \frac{V_g}{\sqrt{R_{gj}^2 + \left(\frac{1}{j\omega C_g}\right)^2}},$$

where $V_g$ is amplitude of gate voltage. The equivalent capacitance of the secondary part is

$$C_r = \frac{C_g C_2}{C_2 + C_g}.$$  

The power-conversion efficiency is maximized when the resonant circuit at the secondary part is resonated with operating frequency $f$ [11], namely

$$C_r = \frac{1}{\omega^2 L_2 C_g}.$$  

Therefore, the secondary resonant capacitance is fixed as

$$C_2 = \frac{C_g}{\omega^2 L_2 C_g - 1}.$$  

The loaded quality factor of the resonant circuit is defined as

$$Q_2 = \frac{\omega L_2}{R_g},$$

which is used for obtaining the sinusoidal output.

3.2. Coupling Part

From $C_r$ and $R_g$, the impedance $Z_2$, which is defined in Fig. 1(d), is expressed as

$$Z_2 = r_L + R_g.$$  

From (8), the amplitude of induced voltage from the primary part can be obtained as

$$V_{ind} = |Z_2| I_2 = (r_L + R_g) I_2.$$  

Figure 2: waveforms of proposed resonant gate driver.
In this analysis, the equivalent transformer model in Fig. 1(c) is adopted, where $Z_{sec}$ is defined as the impedance as shown in Fig. 1(d), namely,

$$Z_{sec} = r_L + R_s + \frac{1}{j\omega C_r}. \quad (10)$$

Dotted-line part in Fig. 1(c) is transformed into equivalent resistance and inductance connected in series as shown in Fig. 1(e). From (10), the impedance of $Z_{eq}$ in Fig. 1(c) is

$$Z_{eq} = \frac{\omega^2k^2L_1L_2Z_2}{Z_2^2 + (\omega L_2 - \frac{1}{\omega C_1})^2} + r_L,$$

$$+ j\omega \frac{k^2 L_1 \left[ \frac{L_2 - \frac{V_s}{L_2} + \frac{1}{\omega C_1}}{Z_2^2 + (\omega L_2 - \frac{1}{\omega C_1})^2} + L_1(1 - k^2) \right]}{R_s + r_L + j\omega L_1} \quad (11)$$

Therefore, the equivalent resistance $R_{eq}$ and the equivalent inductance $L_{eq}$ are

$$R_{eq} = \frac{k^2 \omega L_1 L_2}{R_s + r_L + r_i}, \quad (12)$$

and

$$L_{eq} = L_1. \quad (13)$$

### 3.3. Primary part

From the above analysis, the equivalent circuit of the proposed circuit is illustrated as shown in Fig. 1(e). The current through the primary coil $I_1$, which is the amplitude of $i_1$, can be calculated from

$$I_1 = \frac{V_{ind}}{\omega k \sqrt{L_1 L_2}} = \frac{(r_L + R_s)I_2}{\omega k \sqrt{L_1 L_2}}. \quad (14)$$

For achieving the specified drive voltage and the class-E ZVS/ZDS conditions at turn-on instants, $Z_{eq}$ is transformed into $Z_{eq} = R_{eq} + jL_{eq}$ by the impedance transformation capacitance $C_p$ as shown in Fig. 1(f). $R_{inv}$ and $L_{inv}$ are obtained as [11]

$$R_{inv} = \frac{R_{eq}}{\omega^2 C_p^2 \left[ R_{eq}^2 + (\omega L_{eq} - \frac{1}{\omega C_p})^2 \right]^2}, \quad (15)$$

and

$$L_{inv} = \frac{L_{eq}(1 - \omega^2 L_{eq} C_p) - C_p R_{eq}}{2 \omega^2 C_p^2 \left[ R_{eq}^2 + (\omega L_{eq} - \frac{1}{\omega C_p})^2 \right]^2}, \quad (16)$$

respectively, where $D$ is the on-duty ratio of switch $S$, $\phi_{inv}$ is the phase shifts between the driving signal $D$, and the inverter current $i_{inv}$. $\phi_{inv}$ is [12],

$$\phi_{inv} = \pi + \tan^{-1} \frac{\cos(2\pi D) - 1}{2\pi(1 - D) + \sin(2\pi D)}. \quad (17)$$

$L_{inv}$ is divided into $L_0$ and $L_s$, where $L_0$ and $C_1$ realize resonant circuit for operating frequency $f$. Therefore, $L_s$, which is used for current phase shift for achieving the class-E ZVS/ZDS conditions, can be obtained as

$$\omega L_s \frac{R_{inv}}{L_{inv}} = \frac{2(1 - D)^2 \pi^2 - 1 + 2 \cos \phi_{inv} \cos(2\pi D + \phi_{inv})}{4 \sin(\pi D) \cos(\pi D) \sin(\pi D + \phi_{inv})}$$

$$+ \cos 2(\pi D + \phi_{inv} \cos(2\pi D) - \pi(1 - D) \sin(2\pi D))] + \frac{1}{4 \sin(\pi D) \cos(\pi D) \sin(\pi D + \phi_{inv})} \quad (18)$$

On the other hand, $R_{inv}$ has another expression from the restriction of the class-E ZVS/ZDS conditions as

$$R_{inv} = \frac{2 \sin^2(\pi D) \sin^2(\pi D + \phi_{inv}) V_D^2}{\pi^2(1 - D)^2 R_s^2 (r_s + r_i)}. \quad (19)$$

From (15) and (19), $C_p$ can be obtained as

$$C_p = \frac{\omega L_{eq} R_{inv} \pm \sqrt{R_{inv}^2 [R_{eq}^2 R_{eq}^2 + \omega^2 R_{eq} L_{eq}^2]}}{\omega R_{inv}^2 [R_{eq}^2 + \omega^2 L_{eq}^2]}, \quad (20)$$

It is shown in (20) that there are two candidates of $C_p$. $C_p$ for satisfying $L_0 = L_{inv} - L_s > 0$ should be selected because $L_0$ resonates with $C_1$. $C_1$ and $C_s$ for satisfying the class-E ZVS/ZDS conditions are

$$C_1 = \frac{1}{\omega^2 L_0} = \frac{1}{\omega^2 (L_{inv} - L_s)}, \quad (21)$$

and

$$C_s = \frac{1}{\omega^2 (1 - D) R_{inv}} \left( \frac{2 \sin(\pi D) \cos(\pi D + \phi_{inv})}{\sin(\pi D + \phi_{inv})} \right) \left[ (1 - D) \pi \cos(\pi D) + \sin(\pi D)) \right]. \quad (22)$$

The dc-feed inductance $L_C$ for ensuring less than 10 % current ripple of the input current is expressed as [12]

$$L_C = \frac{R_{inv}}{f} \left[ \frac{2n^2}{4} + 1 \right]. \quad (23)$$

### 4. Design Example

#### 4.1. Design Specifications

In this section, a design example of the proposed driver along with PSpice simulation results is shown. We consider to design a driver for the SiC MOSFET SCT2450KE. The gate capacitance and resistance were measured as $C_g = 3 \, \text{nF}$ and $R_g = 18.75 \, \Omega$. In addition, the amplitude of the gate voltage is specified as $V_g = 15 \, \text{V}$ because of the maximum gate voltage is 20 V. The design specifications of the proposed gate driver were given as operating frequency $f = 7 \, \text{MHz}$, dc-input voltage $V_D = 20 \, \text{V}$, on-duty ratio $D = 0.5$, and loaded quality factor of the secondary resonant circuit $Q_s = 3$. Additionally, the specifications of the coupling coils are necessary. In this paper, loosely coupled coils are used to get leakage flux. We gave the coupling coefficient $k = 0.2$ and the ratio of the number of turns $L_1/L_2 = 1$. 

- 184 -
4.2. Design Procedure

4.2.1. Secondary part design

From $Q_2 = 3$ and specified $C_g$ and $R_g$, the secondary inductance and capacitance are fixed as $L_2 = 1.28 \, \mu\text{H}$ and $C_2 = 467 \, \text{pF}$, respectively. Therefore, we have $L_1 = 1.28 \, \mu\text{H}$ from $L_1/L_2 = 1$.

From $C_g = 3 \, \text{nF}$ and $C_2 = 467 \, \text{pF}$, we have $C_r = 404 \, \text{pF}$. In addition, the amplitude of the secondary current can be obtained from (3) as $I_2 = 0.742 \, \text{A}$. Therefore, the amplitude of the input voltage of the rectifier should be $V_{\text{ind}} = 14.5 \, \text{V}$ from (9). The equivalent inductance and resistance of the secondary part is obtained from (12) and (13) as $R_{eq} = 6.46 \, \Omega$ and $L_{eq} = 1.28 \, \mu\text{H}$, respectively.

4.2.2. Primary and coupling part designs

From (14), the amplitude of the output current of the inverter is $I_1 = 1.29 \, \text{A}$. From the above obtained values and (20), we have $C_p = 228 \, \text{pF}$. For satisfying the class-E ZVS/ZDS conditions, the capacitances should be $C_1 = 287 \, \text{pF}$ and $C_2 = 111 \, \text{pF}$ from (21) and (22), respectively. Finally, the dc-feed inductance for ensuring less than 10% current ripple of the input current is expressed as $L_C > 37.3 \, \mu\text{H}$. $L_C = 40 \, \mu\text{H}$ is used in the simulation.

4.3. Experiment and PSpice-Simulation Results

Figure 3 shows waveforms of the designed driver obtained from analytical expressions, PSpice simulation, and circuit experiment. It can be confirmed from Fig. 3 that the switch voltage of all results achieved the class-E ZVS conditions and the specified amplitude of the drive voltage. In addition, it is also confirmed that the simulation and experimental waveforms agreed with the analytical waveforms quantitatively. It can be stated from these results that the analytical expressions and design procedure have sufficient validity for designs of the proposed driver.

5. Conclusion

This paper has presented the isolated class-E driver for SiC MOSFETs along with its analysis and design procedure. The proposed driver achieves high power-conversion efficiency at high frequencies. The validity of the analytical expressions and design procedure were shown by the quantitative agreements of waveforms among analytical expressions, PSpice-simulation, and circuit experiment.

Acknowledgments

A part of this research was supported by the SIP next-generation power electronics project (NEDO).

References


Flyback converter using SiC power-MOSFET to achieve high frequency operation over 10 MHz

Nobuo Satoh†, Yasuyuki Nishida‡

Department of Electrical and Electronic Engineering, Chiba Institute of Technology
2-17-1 Tsudanuma, Narashino, Chiba, 275-0016 Japan
E-mail: †satoh.nobuo@it-chiba.ac.jp, ‡yasuyuki.nishida@it-chiba.ac.jp

Abstract—We have developed the isolated flyback converter which can be operated at switching frequencies beyond 10 MHz. We estimated the high frequency characteristics of passive devices in various places of the converter circuit. The flyback converter has been examined by means of very high switching frequency of 20 W. In order to verify the fast switching operations, the input and output voltages with currents are measured, the transient behavior of the converter without/with snubber were experimentally estimated both in current continuous and discontinuous operation.

1. Introduction

Flyback converters also face to the electrical insulation used in the majority of small power switching supplies [1, 2]. It is required in the regulations that a power supply ensures the electric insulation to avoid damage to precision instruments, to prevent electric shock to the human body, and to intercept conduction noise. The insulation device in the configuration is a pulse transformer as the choke coil. Plainly, the gate driving circuit must be insulated with keeping detachment of control and power circuits.

In DC power supplies, the switching operation with high-speed can decrease the quantity of energy treated in one cycle operation [3]. This promotes the downsizing of the passive elements of the inductor and the capacitor. However, parasitic impedances in the circuit decide high frequency with the high-speed switching operation. The switching loss occurs because of the surge voltage, the tail current, and ringing phenomena originated in both circuit patterns and switching device parasitic components. Moreover, the destruction of circuits and devices is eventually induced by the heat generated by losses. Thus, one of our goals is aimed at the suppression of losses.

In this study, we showed a flyback converter of 20 W output-class without/with a snubber circuit at 1 MHz switching operation. The transient behavior of the converter was experimentally estimated both in current continuous and discontinuous operation. It described for the experimental study aimed at achieving the 10 MHz switching operation with an output-power of tens of watts [4].

2. Experimental setup

The flyback converter under test is presented in Figure 1. We designed for high-power output as 20 W at 1 MHz operation.

Figure 1: (a) Schematic circuit diagram of the flyback converter using the isolated gate driver and the transformer. (b) Photograph of the flyback converter

In the primary side of the pulse transformer, corresponding to the power input side, the electric power supply (E) is connected. As an important matter, to examine the circuit behavior, there were need to suppress inadvertent oscillation phenomenon. Thence, the prepared power supply is a series-regulator type for stabilizing DC power source (GPO110-3; Takasago Ltd.) rather than the switching type of general-purpose one. In order to manage the maximum attainment voltage, a capacitor (C1) was connected with a charge capacity of 82 µF (withstand voltage: 400 V; Nippon Chemi-Con Corp.) to avoid the instability of voltage supplied at the primary side.

The secondary side of the transformer is connected to a
load at power output side. For high-speed switching operation, the capacitor ($C_2$) is worked as the smoothing capacitor on the output side. A multilayer ceramic capacitor of surface mounts type has been selected as a candidate for the high-frequency application. It is preferable rather than an electrolytic capacitor [5]. The reason is the neglected lifetime deterioration caused by heat, the smaller spatial volume, the high permissible ripple electric current, and the small equivalent series resistance (ESR). These features are strongly expected to raise significant advantages to the miniaturization of the power supply circuit with the high-speed switching. Based on the switching frequency ($f_{sw}$) and the accepted ripple rate, ceramic capacitors are connected with ten-pieces in parallel of surface-mount type as 0.1 $\mu$F (withstand voltage: 50 V; Murata Manufacturing Co., Ltd). Non-inductive load resistor 5 $\Omega$ was terminated as a resistance value to the output.

2.1. Switching device

We have selected the SiC-MOSFET as a switching device, which are evaluated in the flyback converter. There is expected to perform at a high withstand voltage and a large current, being applied in the switching device [6]. The SiC MOSFET (SCTMU001F; Rohm) is a power MOSFET made of a silicon carbide semiconductor with two different implanted regions - p-body region and n-source region - to provide high voltage and the large current capabilities. The resistance and capacitor ($RC$) snubber circuit connected with the power-MOSFET in parallel was inserted for the suppression of the surge voltage. The circuit parameter of $RC$ snubber calculated referring to the guideline for the circuit design by reference [7], and employed such as $R=100 \Omega$, $C=800$ pF.

2.2. Gate driver

Frequency characteristics of the isolated gate-driver that functions as the controlling of electric potential level for the power MOSFET in the flyback converter are shown in Fig. 2. In this case, the band width is approximately 30 MHz. We have confirmed that the consideration of the phase delay is necessary for the phase margin in a switching operation that exceeds 10 MHz.

2.3. Transformer

In the flyback converter, a secondary coil of the pulse transformer is employed as a choking coil [1]. It is also well known that power loss at high-speed switching operation limits the use of magnetic components [8]. A transformer for the flyback converter with high-frequency operation has been designed. The frequency characteristics of the transformer with toroidal-winding and Liqualloy$^\text{TM}$ material has been appraised.

Figure 3 shows the frequency response of the transformer as primary winding and secondary winding, respec-
tively. The inductance and coupling coefficient of the transformer were measured at an operation frequency which equals 10 MHz. The inductance of primary side was 20 μH and the secondary side 2 μH. The coupling factor was estimated at 0.98 by an estimated-circuit of open-short-test. In general, a coupling coefficient evaluated more or equal to 0.95 is required for the transformer [1].

2.4. Rectifying device

Firstly, a SiC Schottky barrier diode (SBD) was adapted to the rectifying device (D). However, in the above operating frequency 1 MHz, the flyback converter circuit was not a stable operation mode. One of the reasons was due to the frequency dependence of the rectifying characteristics of the diode (D). In the 1 MHz operation of the diode as shown in the Fig. 4, it does not have sufficient rectifying characteristics.

![Figure 4: Rectifying properties of diodes according to the operating frequency.](image)

Consequently, we utilized the two of the power MOS-FET, which is a switching device. The rectification in the secondary side was realized by the phase modification.

3. Results and Discussions

All signals were simultaneously observed by two insulation oscilloscopes (TDS2024B; Tektronix Inc.) and two current probes (TCP305; Tektronix Inc.) while analyzing the transient behavior of the flyback converter. The measured states variables were the following: the gate-source voltage (v_{gs}), the drain-source voltage (v_{ds}) of the power MOSFETs, the input voltage (v_{in}), the current of the primary side (i_1), the output voltage (v_{out}), and the current of the secondary side (i_2).

The switching trigger signal is transmitted to the isolated gate driver through the pulse generator (81101A; Agilent). The duty ratio was set within the range of 20% to 50% in order to avoid the continuous-mode operation of the flyback converter. To achieve and evaluate the fast switching operation for all power MOSFETs, the gate-source voltage (v_{gs}) were set at the pulse-shaped waveform with the binary level at 0 V (on-state) and 12 V (off-state).

3.1. Without RC snubber

The waveforms of the gate-source voltage (v_{gs}) and the drain-source voltage (v_{ds}) at f_{sw} = 1 MHz switching operation is shown in Figure 5. Here, The DC voltage of the input-side was 90 V applied. The drain-source voltage (v_{ds}) waveform shows the surge voltage reached at approximately 500 V of its maximum. The discharge current from the choking coil was observed by the secondary current (i_2), which caused ringing phenomena. Nevertheless, the behavior of the flyback converter was scrutinized reliably. The output provides the direct voltage 10 V and the direct current of 2.0 A.

![Figure 5: Oscillograms of v_{gs}, v_{ds}, v_{in}, v_{out}, i_1 and i_2 for the flyback converter without RC snubber at f_{sw}=1 MHz in 20 W output.](image)

3.2. With RC snubber

Figure 6 shows the waveforms of voltage and current that we designed the flyback converter with the snubber circuit at 1 MHz switching operation in 20 W output-power, the output supplies the direct voltage 10 V and the direct current of 2.0 A. The drain-source voltage (v_{ds}) waveform shows the surge voltage reached at approximately 300 V of its maximum. The surge voltage was suppressed for certain, it has been confirmed experimentally. In addition, the discharge current from the choking coil was observed by the secondary current (i_2), which not caused ringing phenomena.

On the other hands, the DC voltage of the input-side was 100 V applied. Apparently the loss was increased by that they have installed the RC snubber circuit. It was confirmed that remain, transient behaviors of the secondary current (i_2) correspond to the discharge current from the choking coil according to the switching operation.
4. Actual and future tasks

We had clarified that there is a limit of 3 MHz operation, experimentally. Its main cause might be the parasitic capacitance of the switching device. We verify at 10 MHz operation of flyback converter, by switching-device having a small parasitic input capacitance is employed.

Moreover, we attempt the suppression of the switching losses by the self-induced soft switching operation due to the ringing frequency of the secondary current, that had shown in the flyback converter without snubber during the switching operation at $f_{sw}=1$ MHz.

5. Summary

We have fabricated the flyback converter that has power output of the 20-watt class at $f_{sw}=1$ MHz. Based on evaluating the device performance of the circuit, the rectification by the switching device is carried out instead of a diode. The circuit was able to operate while ensuring the electrical insulation by the isolated-gate-driver and the transformer.

Acknowledgments

This research was partially supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) for Next-generation power electronics, NEDO.

References


### A three-variable ultralow-power analog silicon neuron circuit

Takashi Kohno\(^\dagger\) and Kazuyuki Aihara\(^\ddagger\)

\(^\dagger\)Institute of Industrial Science, University of Tokyo
4–6–1 Komaba, Meguro-ku, Tokyo 143-8505, Japan
Email: kohno@sat.t.u-tokyo.ac.jp, aihara@sat.t.u-tokyo.ac.jp

\(^\ddagger\)Institute of Industrial Science, University of Tokyo
4–6–1 Komaba, Meguro-ku, Tokyo 143-8505, Japan
Email: kohno@sat.t.u-tokyo.ac.jp, aihara@sat.t.u-tokyo.ac.jp

#### Abstract

A silicon neuronal network is a most fine granular approach to the neuromorphic systems whose significance is growing as a candidate for the core technology of the next generation low-power, autonomous, and intelligent computing systems. In silicon neuron circuits, there has been a trade-off between the power consumption and the capability of reproducing complex neuronal activities. We developed an ultralow-power silicon neuron circuit that can realize multiple classes of neuronal activities including square-wave bursting. Simulation results of our circuit in a square-wave bursting setting are reported.

#### 1. Introduction

The growing social demand for huge-scale information networks and the environmental demand for reducing the power consumption of computing systems are enhancing the significance of neuromorphic systems which realize a low-power, autonomous, and robust computing by mimicking the information processing of the brain. A silicon neuronal network is a most fine granular approach to the neuromorphic systems, which aims to realize an electronic-circuit version of the nervous system by connecting silicon neuron circuits via silicon synapse circuits.

The electrophysiological activity of the neuronal cells is one of the crucial factors in the information processing of the nervous system. Therefore, various models with different levels of details have been developed. The ionic conductance models which describe the mechanisms for the dynamical change of the membrane potential can precisely reproduce complex neuronal activities. Silicon neuron circuits for bio-silico hybrid systems implement them\(^1\) and inherit this advantage. But the complexity in their equations rises their power consumption beyond 100 \(\mu\text{W}\). For low-power consuming systems, simpler models have to be adopted. Many low-power silicon neuron circuits implement the integrate-and-fire(I&F)-based models and consume just several nanowatts\(^2\). The equations of these models are generally simple because they treat the neuronal spike as an event and describe only its timing. It is shown that networks of the I&F-based silicon neurons can execute various processing similar to those by the artificial neural networks\(^3\).

In the brain, however, the neuronal spikes are not uniform\(^4\) (graded response) and chemical synapses can transmit their variation to their postsynaptic cells\(^5\). Thus, there is a possibility that the analog information of neuronal spikes is playing some roles in the information processing of the brain. Another class of simple neuronal models that do not ignore the spike generation mechanisms is the qualitative models. They describe the dynamical structures in the neuronal activities by relatively simple polynomial-based equations. In our previous works\(^6\), we proposed to design an implementation-oriented qualitative neuronal model that is described by formulae of the input-output characteristics of low-power circuits. On the basis of this approach, we developed a three-variable qualitative silicon neuron model for implementation by subthreshold metal-oxide-semiconductor field-effect transistor (MOSFET) analog circuits. It supports multiple classes of neuronal activities including the Class I and II in Hodgkin’s classification, the regular spiking, the square-wave bursting (SWB), and the elliptic bursting by appropriate configurations of its parameters. A fabricated circuit could realize all of these activities with power consumption less than 72 \(\text{nW}\). This power consumption is more than one order of magnitude higher than that of the ultralow-power circuits in\(^7\) (2 \(\text{nW}\)) which was designed by a similar approach but dedicated only to simplest activities, either Class I or Class II.

Because a wide variety of complex neuronal activities are observed in the brain, it is natural to suppose that the capability of reproducing more complex activities is important for the brain-like computing. Hence, to reduce the power consumption penalty paid for this capability, we designed another three-variable model for ultralow-power circuitry by expanding the two-variable model of the silicon neuron circuit in\(^8\) that supports only the Class I and II. In this article, we report simulation results of its implementation in a SWB setting. The model and circuitry of our circuit are explained in the next section. The simulation results are reported in the third section followed by conclusion.

#### 2. Model and Circuitry

Our silicon neuron model has three variables, \(v, n, \) and \(q\). The first two variables are for the spike generation dynamics and the last variable provides slow feedback dynamics that modifies the former faster dynamics. Its equations are

\[
C_v \frac{dv}{dt} = f(v) - g(v) + I_{av} - r(n) - r(q) + I_{stim}, \quad (1)
\]
pacitances \( C_v \), \( C_n \), and \( C_q \) are 0.6 pF, 0.9 pF, and 24 pF, respectively. Currents \( I_{av}, I_{am}, \) and \( I_{aq} \) are parameters and \( I_{stim} \) is a stimulus input. Functions \( f_0(v), g_a(v) (x = v, n), \) and \( r_q(y) (y = n, q) \) are the formulae of the input-output characteristics of transconductance circuits whose schematic is illustrated in Fig. 1. Their equations are

\[
f_0(v) = \frac{M_x}{1 + \exp(-\frac{v - \delta_x}{\kappa})},
\]

\[
g_a(v) = I_0 \sqrt{\frac{R_{201} \exp\left(\frac{-v - \theta_{gx}}{R_{201}}\right)}{1 + R_{201} \exp\left(-\frac{-v - \theta_{gx}}{R_{201}}\right)}},
\]

\[
r_q(y) = I_0 \sqrt{\frac{\exp\left(\frac{v - \theta_{rx}}{R_{201}}\right)}{1 + \exp\left(-\frac{v - \theta_{rx}}{R_{201}}\right)}}.
\]

where \( \kappa, U_T, \) and \( I_0 \) are the capacitive-coupling ratio, the thermal voltage, and the current scaling parameter of PMOS transistors, respectively. Parameter \( \theta \) depends on voltage \( V_{in} \) in Fig. 1(b). Parameters \( R_{201} \) and \( R_{201} \) are controlled by digital inputs bSW0 and bSW1 (see Fig. 1(b)), respectively. When bSW0 (bSW1) is low, the gate voltage of M4 (M5) is shorted to VddSW which equals to or is higher than Vdd. In this situation, M4 (M5) is disabled and thus \( R_{201} \) (\( R_{201} \)) is 1. When voltage bSW0 (bSW1) is high, M4 (M5) is activated and thus \( R_{201} \) (\( R_{201} \)) is 2 (1). These parameters are used to change the \( v \)-offset of \( g_a(v) \). Because the circuit in the gray box is not implemented for \( g_a(v) \), \( R_{201} \) and \( R_{201} \) are fixed to 2. These functions have a monotonic increasing sigmoidal shape and the square root in \( g_a(v) \) and \( r_q(y) \) makes their gradient shallower than that of \( f_0(v) \). The reversed \( N \)-shaped \( v \)-nullcline, which is common in neuronal spike generation systems, is realized by combination of a shallow sigmoidal curve and a steep one.

Figure 2 illustrates the block diagram of our circuit. Each colored box represents the elemental circuit that corresponds to the function in its label. Constant current sources for \( I_{av} \) are implemented by transconductance amplifiers. Each variable in the model is coded by the voltage difference between Vdd (1.0 V) and the non-grounded terminal of the corresponding capacitor. The white box at the top of the diagram is a feedback amplifier for voltage clamp measurement. It clamps \( v \) at a voltage similar to \( V_c \) when SW1 and SW2 are switched to the upper terminal and \( V_{eq} \) is fixed at 1.0 V to keep the output of the upper \( r_q(q) \) circuit zero (the voltage clamp mode). The nullcline of each variable is drawn by plotting the voltage that codes the variable while slowly sweeping \( V_c \). A bifurcation diagram of the spike generation system can be drawn when SW1 is switched to the lower terminal and SW2 is switched to the upper terminal (the bifurcation diagram mode). In this setting, the circuit for the \( v-n \) system operates normally while the output of the \( r_q(q) \) circuit can be controlled by \( V_{eq} \). By slowly sweeping \( V_{eq} \) and plotting the stable state of the \( v-n \) system, we can draw a bifurcation diagram whose bifurcation parameter is \( q \). These two modes guide the parameter voltage tuning process to construct the dynamical structure.
in a neuronal activity to be realized. They are particularly effective to find appropriate parameter voltages of a fabricated circuit which is affected by the device mismatch.

3. Simulation Results

We designed our silicon neuron circuit using a Taiwan Semiconductor Manufacturing Company (TSMC) 0.25 μm mixed-signal CMOS process development kit and found appropriate parameter voltages for a SWB in the Spectre simulation. Figure 3 shows the v- and n-nullclines drawn in the voltage clamp mode. They are configured to have one intersection when q is 0 and three intersections if the v-nullcline is displaced downward when q is increased. To invoke a saddle-loop bifurcation after this saddle-node bifurcation, the time constant of n is suppressed by keeping \( V_m \) in the \( r_s(n) \) circuit sufficiently high. The stable state of the v-n system is projected on the v-q plane in Fig. 4, which is obtained in the bifurcation diagram mode. The red (blue) curve was obtained by slowly increasing (decreasing) \( V_{eq} \). Bistability between a limit cycle and an equilibrium is seen at the region where \( q \) is between 50 mV and 120 mV. The equilibrium point and the limit cycle correspond to the non-spiking and repetitively spiking states, respectively. The sweep time was 2 sec for this diagram. In the increasing sweep, the period of the limit cycle is extended up to infinity as \( q \) is increased. Thus, the sweep time has to be extended depending on the required precision in the diagram. The green curve is the q-nullcline obtained in the voltage clamp mode, which is configured to separate these concurrent two stable states. Note that \( \frac{dq}{dt} \) is positive (negative) above (below) the q-nullcline. When the state point is attracted by the equilibrium, it slowly moves leftward until the equilibrium vanishes at near \( q = 50 \) mV (the silent phase). Then, the state point is attracted by the limit cycle (the spiking phase). It slowly moves rightward until the limit cycle vanishes at near \( q = 120 \) mV and attracted by the equilibrium again. This alternation of the silent and spiking phases is the mechanism of the SWB. We observed a regular bursting (Fig. 5(a)) similar to that in the SWB cells[18, 19]. It is known that the tonic firing and the chaotic bursting are observed in the Hindmarsh-Rose model, a qualitative SWB model, when a parameter for the slowest variable is modified[20]. Our circuit could produce these activities (Figs. 5(b) and (c)) by appropriately tuning a parameter voltage that controls \( M_q \). In all these settings, the average power consumption was less than 4.9 nW.

4. Conclusion

We designed a configurable ultralow-power silicon neuron circuit and verified that it can realize SWB using the Spectre simulator. It has 14 parameter voltages which have to be configured appropriately. For the configuration pro-
Table 1: Comparison of analog silicon neurons (*exact value not shown, **graded response*)

<table>
<thead>
<tr>
<th></th>
<th>this work</th>
<th>[17]</th>
<th>[4]</th>
<th>[7]</th>
</tr>
</thead>
<tbody>
<tr>
<td>model</td>
<td>QM</td>
<td>QM</td>
<td>I&amp;F</td>
<td>I&amp;F</td>
</tr>
<tr>
<td>power</td>
<td>&lt; 5 nW</td>
<td>1.7 nW</td>
<td>&lt; 3 nW**</td>
<td>&lt; 7 nW</td>
</tr>
<tr>
<td>area</td>
<td>90000</td>
<td>2740</td>
<td>918</td>
<td>2980</td>
</tr>
<tr>
<td>process</td>
<td>0.25 µm</td>
<td>0.35 µm</td>
<td>0.18 µm</td>
<td>90 nm</td>
</tr>
<tr>
<td>Class I</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>partially</td>
</tr>
<tr>
<td>Class II</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>GR**</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>SWB</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
</tbody>
</table>

procedure, the feedback amplifier and switches are integrated in our circuit. They are used to draw the nullclines and the bifurcation diagrams of the v-r system, which provide an effective guide to construct proper dynamical structures. Its power consumption was estimated to be lower than 5 nW, which will be reduced in our future circuits by substituting power parameter voltage generator circuits which consume 1001 pA, 929 pA, and 313 pA to generate 173 pA, 156 pA, and -32 pA, respectively. The power consumption will be reduced to about 3 nW when the sum of the differences between these currents is saved by this substitution. We are working on ultralow-power parameter voltage generator circuits which consume about 100 pW. If they are integrated into our silicon neuron circuit, the total power consumption will be about 4.4 nW. The difference between it and 1.7 nW in [17] will be the cost for the configurability and the capability of realizing complex neuronal activities. Our circuit occupies larger area than recent analog silicon neurons (see Table 1). However, it may be shrunk by using a finer process because the effects of device mismatch are compensated by our parameter tuning procedure.

Acknowledgments

This study was partially supported by JST PRESTO and CREST program, VLSI Design and Education Center(VDEC), The University of Tokyo with the collaboration with Cadence Corporation.

References

Two Heuristic Approaches to Parameter Tuning for an Analog Silicon Neuron Circuit

Ethan Green† and Takashi Kohno†
† Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro, Tokyo, 153-8505, Japan
Email: green@sat.t.u-tokyo.ac.jp, kohno@sat.t.u-tokyo.ac.jp

Abstract – Neurons play a major role in memory, cognition, sensory processing, body regulation, and a host of other functions vital to organisms. Analog silicon neurons are biologically inspired VLSI (very-large-scale integrated) circuits that mimic the electrophysiological behavior of neurons. This research looks at circuit parameter tuning for an ultra-low power analog silicon neuron designed with qualitative neuronal modeling. A key challenge to operating this circuit is adjustment of the circuit parameters to allow for similar behavior across a range of temperatures and eventually amongst many silicon neuron circuits in a silicon neuronal network. Two heuristic approaches were applied to the silicon neuron to supplement trial-and-error-based tuning of the circuit’s parameter voltages. In the future, these two approaches will be combined to create a fully automated tuning algorithm.

1. Introduction

Analog silicon neurons are electronic circuits that adopt the electrophysiological characteristics of neurons, the principle cells of the nervous system. These circuits operate in continuous time, require minimal power, and can be integrated into massively parallel networks [1]. These silicon neuronal networks may form the basis of future computers with neuromimetic architecture that may supplement digital transistor logic with novel computing techniques. Future uses of silicon neuronal networks may include autonomous machines, bio-silico hybrid devices, and ultra-low-power computing platforms.

Qualitative neuronal modeling refers to the use of approximation to reduce the complexity of biophysically accurate ionic-conductance models to create simpler mathematical models which maintain similar neuronal dynamics with fewer variables and simpler formulae. The silicon neuron used in this research was designed with a qualitative modeling approach in order to implement simpler circuitry. The silicon neuron can replicate Class I and Class II spiking as defined by Hodgkin [2] with only two variables [3].

While simpler than conductance-based silicon neuron circuits, our silicon neuron still requires many parameter voltages—bias voltages applied to the components comprising the circuit. A challenge to proper operation of the circuit is finding parameter voltages that can satisfy narrowly defined criteria to yield neuron-like operation. The circuit is equipped with feedback amplifiers that guide finding appropriate parameter voltages with a procedure similar to voltage clamp experiments. Since the circuit uses subthreshold-operated MOSFETs, operation suffers from pronounced temperature sensitivity and requires the parameter voltages to be varied with temperature.

In our previous work [4], we showed that trial and error can be used to find parameter voltage sets for different temperatures. We identified pillar sets of parameter voltages at 17, 22, 27, 32, and 37°C and then interpolated parameters for intermediary temperatures and reported how successfully these sets replicated benchmark circuit behavior.

In this work, we discuss an algorithm that can more effectively generate parameter voltages for pillar temperatures and also be used to tune silicon neuron circuits in a future silicon neuronal network. We show the merits and drawbacks of two algorithms, one based on brute force and the other on Differential Evolution, and then propose a future tuning algorithm that combines the strongpoints of these two approaches. The Spectre circuit simulation platform was used for circuit simulations. The model and circuit of our silicon neuron is explained in the next section. The trial and error approach is reviewed in Sec. 3. The two new approaches are reported in Sec. 4, which is followed by a conclusion.

2. Silicon Neuron Model and Circuit

The silicon neuron circuit [3] is divided into two blocks, a $r$-block and $n$-block, each with a capacitor that is charged and discharged by transconductance circuit components (Fig. 1). Variable $r$ represents the membrane potential and variable $n$ represents an abstracted ionic activity. Each variable is coded by subtracting the voltage over its capacitor from $V_{ds}$.

The system equations of the silicon neuron are:

\[
\begin{align*}
C_v \frac{dv}{dt} &= f_v(v) - g_v(v) + I_{av} - r(n) + I_{stim}, \quad (1) \\
C_n \frac{dn}{dt} &= f_n(v) - g_n(v) + I_{an} - r(n), \quad (2)
\end{align*}
\]

where $f_v(v)$, $g_v(v)$, and $r(n)$ are the equations governing the transconductance circuit components, $I_{av}$ is a constant current $(x=v,n)$, and $I_{stim}$ is an externally applied stimulus
current. The $f_x(v)$ circuit is a differential pair tied to a cascaded current mirror, and $g_x(v)$ and $r(n)$ are modified cascode circuits with source degeneration (Fig. 2). The current-voltage equations for these components are:

$$f_x(v) = \frac{M_x}{1 + \exp \left( \frac{-v - \delta_x}{\kappa} \right)}$$  

(3)

$$g_x(v) = I_0 \left[ \frac{1 + \exp \left( \frac{-v - \theta_x}{\kappa} \right)}{1 + \exp \left( \frac{-v - \theta_x}{\kappa} \right)} \right]$$  

(4)

$$r(n) = I_0 \left[ \frac{1 + \exp \left( \frac{-n - \theta_r}{\kappa} \right)}{1 + \exp \left( \frac{-n - \theta_r}{\kappa} \right)} \right]$$  

(5)

where constants $M_x$, $\delta_x$, $\theta_x$ (x=v,n), and $\theta_r$ correspond to externally applied bias voltages (here referred to as parameter voltages), $\kappa$ is the PMOS capacitive coupling ratio, $I_0$ is the PMOS off-current, and $V_T$ is the thermal voltage. Eqs. (3)–(5) express sigmoidal relationships [3][5].

All transistors in the silicon neuron circuit are operated in the subthreshold regime, allowing for desirable exponential current-voltage characteristics and total power consumption as low as 3 nW.

The nullclines, curves on which Eqs. (1) and (2) equal zero, can be used to describe the neuron-like dynamics of the circuit [6]. $TAV$ and $TAN$ are the transconductance amplifiers that measure the nullclines with a voltage clamp technique similar to a DC steady state analysis.

3. Parameter Tuning with Trial and Error

Pronounced temperature sensitivity is a major drawback to operating transistors in the subthreshold regime. Temperature changes of a few degrees can completely disrupt normal operation of the circuit. Furthermore, due to transistor mismatch, individual silicon neurons in a future neuronal network will require a unique set of parameter voltages. These factors illustrate the importance of developing a parameter voltage tuning strategy for proper operation of the silicon neuron.

In [4], we performed Spectre simulations in which we adjusted the parameter voltages of the silicon neuron to yield similar dynamics at different temperatures. We established a benchmark at 27°C in which we recorded the circuit’s response to 5 pA and 10 pA sustained stimuli (Fig. 3), and the threshold current necessary to generate an action potential for a 500 µs pulse stimulus ($I_{th}$). We also plotted the nullclines (Fig. 4). All successive attempts to tune the circuit were judged by these benchmarks. Class I spiking was used because of its readily discernable phase plane with three nullcline intersections (Fig. 4) and simple firing mechanism governed by a saddle-node on invariant circle bifurcation [6].
We then simulated the circuit at 17, 22, 32, and 37°C and adjusted the parameter voltages with trial and error. Five influential parameter voltages were selected with all others held constant. The \( g_d(v) \) circuit was turned off. The nullclines were used to find approximate parameter voltages, which were then tuned until circuit operation matched the benchmark transient behavior as closely as possible.

Table 1 shows the parameter voltages found by trial and error for each temperature. Parameter voltages \( g_v \_V_m \), \( f_n \_V_b \), \( I_{av} \_V_in \), \( I_{an} \_V_in \), and \( r \_V_m \) in the table correspond to constants \( \theta_s \), \( M_n \), \( I_{av} \), \( I_{an} \), and \( \theta_n \) in the circuit equations. The bottom three rows of the table note the transient behavior of each parameter set at its temperature.

4. Heuristic Algorithms

Trial and error allowed us to find adequate parameter voltages for a variety of temperatures, but the method was cumbersome and time consuming. To bring a degree of automation to parameter tuning, we wrote a script which calls Spectre, runs transient or nullcline-drawing simulations, and analyzes results. This script can then be integrated with search algorithms to identify parameter voltages that recreate the circuit behavior.

4.1. Brute Force

The first search algorithm used was a brute force method. Our aim was to fine-tune the results of the trial and error approach by searching for better parameter voltages in the nearby vicinity of parameter space. For each of the five parameter voltages, the brute force algorithm tested 5 possibilities: the original value, \( \pm 0.5 \) mV, and \( \pm 1 \) mV. The circuit was then simulated with all possible parameter combinations. The threshold current \( (I_{th}) \) and responses to 5 pA and 10 pA sustained stimuli were recorded. 5\(^n\)=3125 simulations required about 3.5 hours of calculation time.

The algorithm searches for the minimum of a cost function, in this case the magnitude of the difference between vectors of the observed circuit behavior and benchmark behavior (Eqs. (6) and (7)). Vector \( x \) represents the input circuit parameters, function \( j(x) \) is the vector of behaviors as recorded by the circuit simulator, and vector \( a \) is the benchmark behavior. The threshold index is related to the threshold current.

\[
j(x) = \begin{bmatrix} j(x) \text{ threshold index} \\ 5 \text{ pA stimulus response} \\ 10 \text{ pA stimulus response} \end{bmatrix}, \quad a = \begin{bmatrix} 1.21 \\ 15.87 \\ 37.18 \end{bmatrix}
\]

\[
f(x) = ||j(x) - a||
\]

The brute force algorithm includes a polishing step which uses the simulation results to predict the location of the global minimum between parameter combinations.

Table 1: Results of Trial and Error. Reprinted from [4].

<table>
<thead>
<tr>
<th>Temperature (°C)</th>
<th>17</th>
<th>22</th>
<th>32</th>
<th>37</th>
</tr>
</thead>
<tbody>
<tr>
<td>( g_v _V_m ) (mV)</td>
<td>388</td>
<td>417</td>
<td>432</td>
<td>443</td>
</tr>
<tr>
<td>( f_n _V_b ) (mV)</td>
<td>257</td>
<td>248</td>
<td>237</td>
<td>231</td>
</tr>
<tr>
<td>( I_{av} _V_in ) (mV)</td>
<td>434</td>
<td>450</td>
<td>461</td>
<td>468.5</td>
</tr>
<tr>
<td>( I_{an} _V_in ) (mV)</td>
<td>420</td>
<td>449.5</td>
<td>464.5</td>
<td>475</td>
</tr>
<tr>
<td>( r _V_m ) (mV)</td>
<td>480.5</td>
<td>462.5</td>
<td>445</td>
<td>430</td>
</tr>
</tbody>
</table>

For example, for \( f_n \_V_b \) at 22°C, the algorithm tested 247, 247.5, 248, 248.5, and 249 mV and by interpolation returned 247.4 mV as the global minimum.

Table 2 shows the results of the brute force algorithm. Three circuit behaviors were measured for the four temperature steps. The categories which showed improvement from the original trial and error results are highlighted in gray.

4.2. Differential Evolution

The second search algorithm used was Differential Evolution, an evolutionary algorithm that evolves a population of random solution vectors over many generations to find the global minimum of a cost function [7]. Our algorithm was inspired by a similar algorithm in [8].

Instead of running transient simulations, the algorithm employs the silicon neuron’s nullcline drawing feature. The cost function used was the vector magnitude of the mean absolute error of the \( v- \) and \( n \)-nullclines from the benchmark nullclines (Eq. 8). \( x \) again is a vector of the input parameter voltages, \( v(x) \) and \( n(x) \) represent points on the \( v- \) and \( n \)-nullclines, and \( b_i \) and \( c_i \) are the corresponding points on the benchmark nullclines. The summation is carried out over each step of a DC steady state analysis.
\[ f(\mathbf{x}) \triangleq (\sum_{i=1}^{k} |v_i(\mathbf{x}) - b_i|) + (\sum_{i=1}^{n} |c_i(\mathbf{x}) - e_i|) \]

(8)

The algorithm starts with an initial set of random nullclines and gradually fits them to the benchmark curves by rejecting parameter sets which perform poorly while maintaining and evolving sets which perform well. Figure 5 shows the nullclines returned by the algorithm for 32°C overlaid on the benchmark nullclines. One run of the algorithm required about 3 hours of computing time.

The nullclines are independent of the \( r(n) \) circuit, so the parameter voltage for \( \text{rn}_\text{Vm} \)—the only parameter voltage that controls the \( r(n) \) circuit—was taken from the brute force approach and further tuned by trial and error for 37°C. Table 3 shows these parameter sets along with the results of the transient simulations, which were comparable to the original trial and error approach for 17, 22, and 32°C. The 5 pA stimulus failed to induce firing for 37°C. The measured behaviors that were closer to the benchmark than the trial and error results are highlighted in gray. Fewer categories were closer to the benchmark than for brute force, but the Differential Evolution algorithm was significantly more automated.

5. Conclusion

5.1. Discussion

The brute force algorithm worked well by evaluating the circuit’s transient behavior, but suffered from a limited search range, the expansion of which exponentially increases calculation cost. Effective use of this algorithm necessitates a first step which identifies approximate parameter voltages.

The Differential Evolution algorithm on the other hand starts from a random solution and automatically finds parameter voltages that yield accurate nullclines relatively quickly. However, the results of the transient simulations were poor. While the nullclines give good clues about the dynamics of the circuit, they only partially describe the system. Additionally, the algorithm in its current form does not tune \( r(n) \) because of circuit constraints. A more effective algorithm must tune \( r(n) \) and take transient simulations into account.

5.2. Future Algorithm

We propose a 2-stage heuristic algorithm which combines the merits of Differential Evolution and brute force. The first stage will use Differential Evolution to tune the nullclines of the circuit to best match the benchmark nullclines. The second stage will use brute force to search all nearby parameter combinations for the parameter set which most closely replicates the benchmark transient behavior. The brute force stage may also search a wider range for \( \text{rn}_\text{Vm} \) of \( r(n) \) since this parameter is not incorporated in the first stage.

Figure 5: Nullclines from Differential Evolution for 32°C

Such an algorithm may improve the above results. The same approach may also be used to tune different silicon neurons of the same design which vary due to transistor mismatch.

We plan to first simulate this algorithm with Spectre and eventually implement it with our actual VLSI circuit using lab equipment and software such as LabVIEW.

6. Acknowledgements

This study was supported by JST PRESTO and CREST, and VDEC, the University of Tokyo in collaboration with Cadence Design Systems, Inc.

7. References

Elliptic and parabolic bursting in a digital silicon neuron model

Takuya Nanami1 and Kazuyuki Aihara1 and Takashi Kohno1

1, Graduate School of Engineering, The University of Tokyo
Tokyo, Japan
Email: nanami@sat.t.u-tokyo.ac.jp

Abstract—The digital spiking silicon neuron (DSSN) model is a qualitative model designed to be implemented efficiently using digital arithmetic circuits. In our previous study, we extended this model to support the neuronal activities of four cortical and thalamic neuronal classes. In this paper, we further extended this model to reproduce bursting activities of the ionic-conductance models of elliptic and parabolic bursters.

1. Introduction

Silicon neuronal networks simulate neuronal activity with low power consumption and in high speed. They are thought to be a way to realize an intelligent system comparable to the human brain. They can be implemented by both analog and digital circuits. Analog circuit implementation consumes ultra-low power down to several nanowatts per silicon neuron[1][2][3]; however, it involves the technical hurdles of fabrication mismatch and temperature dependence when constructing a large-scale network. On the other hand, digital circuit implementation does not have this limitation because it is far less sensitive to these factors, though power consumption tends to be higher than in analog circuit implementation. In particular, the field-programmable gate array (FPGA) is popular because of its low cost and high availability. Generally, digital silicon neuronal networks implemented on an FPGA perform calculations at a higher speed than the biological real time [4][5][6][7].

The digital spiking silicon neuron (DSSN) model [8] is a qualitative model designed to be implemented efficiently using digital arithmetic circuits. The model simulates diverse neuronal activities with the fixed-point operation and Euler’s method. Given appropriate parameter sets, this model can reproduce the Class I and II in Hodgkin’’s classification [9] as well as Class I* [10], which is defined by a unique mathematical structure.

In our previous study[11], we extended this model to support the neuronal activities of four cortical and thalamic neuronal classes: namely, regular spiking (RS), fast spiking (FS), intrinsically bursting (IB), and low-threshold spiking (LTS). These classes are described by ionic-conductance models [12] that can accurately reproduce neuronal activities. Moreover, by measuring the $C_V$ and $L_V$ [13] statistics of the spike sequences, we confirmed that DSSN and ionic-conductance models have the same statistical properties in each neuron class.

In this work, we realized the elliptic and parabolic bursters with the DSSN model. For the elliptic bursting, we referred to an ionic conductance model in [14]. Wang observed experimental data from layer V pyramidal neurons in the cat sensory motor cortex and modeled them with ionic conductance type equations. We adopted the three-variable DSSN model and determined an appropriate parameter set to mimic Wang’s ionic conductance model in response to input current with various magnitudes. For the parabolic bursting, we developed a four-variable DSSN model and found a suitable parameter set that replicates the behavior of a model in [15]. The Izhikevich model[16] which is popular in silicon neuronal networks cannot reproduce a parabolic bursting because it requires at least four state variables. The remainder of this paper is organized as follows. Section 2 introduces our neuron model. The simulation result is presented in Section 3. Section 4 summarizes the work and suggests ideas for the future.

2. Method

2.1. DSSN model

The DSSN model [8] is a neuron model that can simulate multiple classes of neuronal activities by Euler’s method with fixed point operation. We adopted the 3-variable DSSN model which is expressed as follows:

\[
\frac{dv}{dt} = \frac{\phi}{\tau} \left( f(v) - n - q + I_0 + I_{\text{stim}} \right),
\]

\[
\frac{dn}{dt} = \frac{1}{\tau} \left( g(v) - n \right),
\]

\[
\frac{dq}{dt} = \frac{1}{\tau} \left( h(v) - q \right),
\]

\[
f(v) = \begin{cases} 
  a_{fn}(v - b_{fn})^2 + c_{fn} & (v < 0) \\
  a_{fp}(v - b_{fp})^2 + c_{fp} & (v \geq 0),
\end{cases}
\]

\[
g(v) = \begin{cases} 
  a_{gn}(v - b_{gn})^2 + c_{gn} & (v < r_g) \\
  a_{gp}(v - b_{gp})^2 + c_{gp} & (v \geq r_g),
\end{cases}
\]

\[
h(v) = \begin{cases} 
  a_{hn}(v - b_{hn})^2 + c_{hn} & (v < r_h) \\
  a_{hp}(v - b_{hp})^2 + c_{hp} & (v \geq r_h),
\end{cases}
\]
where \( v \) corresponds to the membrane potential, and \( n \) and \( q \) are the fast and slow variables, respectively, that abstractly describe the activity of the ion channels. The parameter \( I_0 \) is a bias constant and \( I_{\text{stim}} \) represents the input stimulus. Parameters \( \phi, \epsilon \) and \( \tau \) control the time constants of the variables. Parameters \( r_s, a_s, b_s, \) and \( c_s \), where \( x = f_n, f_p, g_n, g_p, h_n, \) or \( h_p \), are constants that adjust the nullclines of the variables. All of the variables and constants in this qualitative model are purely abstract with no physical units. Most existing qualitative neuronal models replicate the spiking dynamics by a cubic variable term [17][18][19]. Because multiplication consumes significant circuit resources in a digital arithmetic circuit, the DSSN model adopts a piecewise quadratic function so that its numerical integration step includes only one multiplication between variables. The elliptic bursting can be realized using a slow variable, but the parabolic bursting requires at least two slow variables. Then, we extended the DSSN model by modifying Eq.(1) and adding Eq.(8) as follows:

\[
\frac{dv}{dt} = \frac{\phi}{\tau} (f(v) - n - q + u + I_0 + I_{\text{stim}}),
\]

\[
\frac{du}{dt} = \frac{\epsilon}{\tau} (v - v_o - au),
\]

where \( u \) is a slow variable that abstractly describes the activity of depolarizing slow ion channels.

![Bifurcation diagram of the DSSN model corresponding to the elliptic bursting.](image1)

**Figure 1:** Bifurcation diagram of the DSSN model corresponding to the elliptic bursting. The fast subsystem (the \( v - n \) system) exhibits a subcritical Hopf bifurcation. Bistability composed of a stable spiral and a stable limit cycle is seen.

![Bifurcation diagram of the DSSN model corresponding to the parabolic bursting.](image2)

**Figure 2:** Bifurcation diagram of the DSSN model corresponding to the parabolic bursting. The fast subsystem exhibits a saddle-node on invariant circle bifurcation.

![Waveforms of (a) the Wang’s and (b) the DSSN models corresponding to the elliptic bursting.](image3)

**Figure 3:** Waveforms of (a) the Wang’s and (b) the DSSN models corresponding to the elliptic bursting. Both models generate periodic burst firing in response to (top) a weak input, (middle) a medium input, and (bottom) a strong input.

![Statistical properties of Wang’s ionic conductance and DSSN models corresponding to the elliptic bursting.](image4)

**Figure 4:** Statistical properties of Wang’s ionic conductance and DSSN models corresponding to the elliptic bursting. The x-axis corresponds to the duration of a silent phase. The y-axis represents the number of spikes in a spiking phase.

### 2.2. Parameter tuning

To reproduce the elliptic bursting, we first tuned the parameters of the fast subsystem (the \( v - n \) system) to realize the subcritical Hopf bifurcation that has a bistable regime including a stable spiral and a stable limit cycle. Figure 1 shows the bifurcation diagram of the fast subsystem. We conducted bifurcation analysis while varying slow variable \( q \) as the parameter. The bistable area and slow state variable \( q \) cause bursting and resting states. Secondly, we tuned the remaining parameters that control the dynamics of the slow subsystem in order to reproduce the neuronal activity of Wang’s ionic conductance model precisely.

To reproduce the parabolic bursting, the fast subsystem
of the model does not require a bistable regime. Bursting
and resting states are generated by two slow state variables.
We first determined the parameters that control the dynamics
of the fast subsystem. The bifurcation diagram is shown in
Fig.2. The fast subsystem exhibits a saddle-node on in-
variant circle bifurcation. Second, we tuned the remaining
parameters that control the dynamics of the slow subsystem
in order to reproduce the neuronal activity of Plant’s ionic
conductance model precisely.

3. Result

3.1. Elliptic bursting

Figure 3 shows waveforms of Wang’s model and the
DSSN model corresponding to the elliptic bursting in re-
sponse to the input stimulus of several magnitudes. We
could not find a parameter set for the DSSN model that
qualitatively reproduce the activity of the Wang’s model.

With both models the number of spikes in a spiking phase
increases and the period of a silent phase decreases as the
stimulus intensity increases.

To verify the qualitative similarity, we visualized these
characteristics by plotting the duration of a silent phase on
the $x$–axis and the number of spikes in a spiking phase on
the $y$–axis by changing the stimulus intensity (Fig.4). Note
that the duration of a silent phase is the dominant compon-
ent of the bursting period. With both models, the number
of spikes decreases as the bursting period increases and two
inflection points (circles in the figure) are seen. We counted
spikes when the membrane potential exceeds a threshold
value, which was set as $V = -20\text{mV}$ in Wang’s model and
$V=0.1$ in the DSSN model.

3.2. Parabolic bursting

Figure 5 shows waveforms of Plant’s model and the
DSSN model in the parabolic bursting mode in response
to the input stimulus with several magnitudes. With both
models, the number of spikes in a spiking phase
increases and a period of the silent phase decreases as the
stimulus intensity increases.

The transition of the two slow variables is plotted on the
$q - u$ plane in Fig.6. Above the dotted line, the model is
in the firing mode and the trajectory is on the stable limit
cycle. Below the dotted line, the model is in the resting
mode and the trajectory is on the stable node.

We evaluated the qualitative similarity between the be-
havior of the DSSN and Wang’s models by the same plot
as Fig.4. They share the negative slope with three inflection
points (circles in the figure) (Fig.7).

3.3. Device utilization

We compiled the DSSN models for Virtex-7
XC7VX690T FPGA using Xilinx Vivado Design Suite.
Device utilization is listed in Table 1. In the elliptic burst-
ing mode, we used 18-bit signed fixed point with 14-bit
fraction part for all variables. In the parabolic bursting
mode, 24-bit signed fixed point with 20-bit fraction part was required to keep high-precision. The DSP unit was used to calculate $v^2$. Table 1 also lists the resource usage in [20] which implemented a fully-connected network of 1024 neurons on a Virtex-5 xc5vlx330t FPGA. They adopted the Izhikevich model and used single or double precision floating-point operations. The resources listed in the table is for calculation of 1024 neurons. It is clear that their circuit consumes far less resources than ours. The difference between the resource requirements for our circuit and the circuit in [20] is the penalty required to dissolve the limitations in the integrate-and-fire-based models.

4. Conclusion

In this work, we tuned parameters of the three-variable DSSN model for the elliptic bursting and realized qualitatively similar behavior to that of Plant’s model. A slow variable was supplemented to the three-variable DSSN model and its parameters were tuned to realize the parabolic bursting qualitatively similar to that of the Wang’s model. The similarity was verified by measuring the duration of the resting state and the number of spikes in the bursting. The DSSN model was numerically solved by Euler’s method ($t=0.0001s$). Previous studies have applied the DSSN model to the Class I and II in Hodgkin’s classification, the square wave bursting and four cortical and thalamic neuron classes. Here we extended and/or configured the DSSN model to match the elliptic bursting and parabolic bursting that are not listed in the repertoire of the Izhikevich model which is a most popular simplified neuron model. It is not elucidated completely what properties of the neuronal activities are playing the key roles in the information processing in the brain. Our models intend to contribute to the “analysis by thynsesis” approach to this question by pursuing qualitative reproduction of as many characteristics of the neuronal activities as possible. In future work, we will improve the equations and parameters to reproduce the elliptic and parabolic burstmts in the Wang’s and the Plant’s models more precisely. For parameter tuning, some heuristic methods such as differential evolution algorithms may be utilized as in [21].

5. Acknowledgement

This work was partially supported by JSPS Grant-in-Aid for scientific Exploratory Research Number 25240045 and JST PRESTO and CREST.

Table 1: Device utilization

<table>
<thead>
<tr>
<th>Name</th>
<th>Block Sparse</th>
<th>Parameter Sparse</th>
<th>LUTs</th>
<th>DSPs</th>
<th>Resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP</td>
<td>13200(9.6%)</td>
<td>31300(9.0%)</td>
<td>493,000</td>
<td>7824</td>
<td>12543</td>
</tr>
<tr>
<td>LUTs</td>
<td>104.00%</td>
<td>104.00%</td>
<td>10000</td>
<td>16</td>
<td>96</td>
</tr>
</tbody>
</table>

References

Spiking neural network simulation on FPGAs with automatic and intensive pipelining

Taro Kawao† Masato Neishi Tomohiro Okamoto Amir Masoud Gharehbaghi
Takashi Kohno Masahiro Fujita
University of Tokyo, Japan
†Now at Renesas Electronics, Japan

Abstract—There are lots of scientific interests in the behaviors of large spiking neural networks. One way to understand the behaviors is to simulate them as fast as possible with customized hardware, such as FPGA. This paper shows highly pipelined implementations of spiking neural networks on FPGA using a high level synthesis tool. Our accelerator allows 256 neurons to operate at 280 times faster than real time brain operations, which is around 8 times faster than the previous reports on similar directions. It is also designed for a multi-FPGA implementation which can simulate up to 3,000 neurons.

1. Introduction

There have been a series of researches on the use of specialized and customized hardware for speeding up particular scientific computations. Completely customized hardware, such as ASIC (Application Specific Integrated Circuit) implemented in silicon does not have any flexibility after fabrication, and there is no way to modify operations even if the target computations need to be changed. On the other hand, programmable and reconfigurable hardware, such as FPGA (Field Programmable Gate Arrays), can change their functionality at anytime, and so it is one of the ideal devices which realize the customized computing machines.

In this paper we show highly pipelined implementation of spiking neural network simulation on FPGA. Neural networks are mathematical models on the functionality of neuron and synapse, and consist of nodes for neurons and edges for synapses. By changing the weights for edges based on learning, various information processing can be realized. By using sufficiently large networks, robust computing, such as computations under various noises, can be achieved for pattern recognition, data classification, and others. There have been developed many FPGA-based spiking neuronal network simulators [1][3][7] [6][2][4].

Here, we use DSSN network model [5] which uses relatively simple computations for ease of hardware implementations but reproduces the behaviors of biological neural networks with sufficient accuracy. Our work is based on the previous works [6][7] but improves the simulation speed by 8 times and also can deal with larger networks. The previous works [6][7] realize the same speed as human brain up to around 1,500 neurons. The scope of our research covers the accelerated simulation of neuronal networks composed of tens of thousands of neurons. This number is comparable to that of the insect brains which realize intelligent and adaptive processing. Thus, speeding up simulations of that size is practically very important.

In our framework, the highly pipelined implementations of spiking neural network simulations are generated through the use of a high level synthesis tool, called Max Compiler [8]. High level synthesis tools can automatically refine designs in high level, such as in C or Java, into the ones in RTL (Register Transfer Level). Once designs in RTL are obtained, there are established tool flows by which FPGA implementations can be automatically obtained. Max Compiler can introduce intensive pipelined designs utilizing their special IPs (pre-designed functional units). The numbers of pipeline stages can be more than 1,000, which means more than 1,000 data parallel operations are processed simultaneously. With the use of high level synthesis tools, high performance neural network simulation on FPGA can be relatively easily realized.

The rest of the paper is organized as follows. In the next section the neural network model we are going to implement is reviewed. Then the previous works are briefly examined in the following section. Our proposed implementation is presented next with its experimental results. The final section gives concluding remarks.

2. Target neural network model

In our target model, there are N neurons which are connected to one another through synapse networks as shown in Figure 1. All neurons are connected to all the other neurons. For each neuron, the input from synapses to the neuron is computed as shown in the expression 5. The computations inside a neuron is shown in the expressions 1, 2, and 3. The output from a neuron to synapses is computed as shown in the expression 4. The learning on the spiking neural network is defined in the expressions 6 and 7. Some of the expressions originally defined as differential equations are converted into the corresponding difference equations.

Here we implement the spiking neural network model called DSSN model [5]. A spike is defined as a narrow
impulse with some amount of voltage swing. The variables \( v, n, \) and \( q \) are introduced for the modeling. By manipulating the variables, \( v \) and \( n, \) spikes are generated, and the variable, \( q \) adjusts the intervals of spikes. \( I_0 \) is a constant and \( I_{sim} \) is the weighted sum of synaptic inputs to a neuron and computed as shown in the expression 5. While \( v \) is above a threshold, \( [T] = 1 \) and else \( [T] = 0. \) \( \varphi \) and \( \tau \) are time constants, and the others, \( v_0, \alpha, a_s, a_p, b_s, b_p, c_s, c_p, k_s, k_p, I_a, I_p, m_a, m_p, \) are control parameters for the behaviors of spikes. For a neuron, \( I_{sim} \) is the input and \( v \) is the output.

\[
\begin{align*}
    v(t + \Delta t) &= v(t) + \Delta t \frac{\varphi}{\tau} \\
    n(t + \Delta t) &= n(t) + \Delta t \frac{1}{\tau} \\
    q(t + \Delta t) &= q(t) + \Delta t \frac{\varphi}{\tau} (v - v_0 - aq)
\end{align*}
\]

The model of synapse is based on [9], and it is defined with the expressions, 4 and 5.

\[
I_s(t + dt) = I_s + \Delta t \times \begin{cases} 
    \alpha (1 - I_s(t)) & ([T] = 1) \\
    -\beta I_s(t) & ([T] = 0)
\end{cases}
\]

\[
I_{sim}^{'} = c \sum_{j=1}^{N} W_{ij} I_j
\]

As for more details, please refer to [5].

### 2.1. Hebbian learning

The Hebbian learning is a learning method based on timings of spikes. When neuron \( i \) and \( j \) generate spikes in a similar timing, neuron \( j \)'s influence on neuron \( i, W_{ij}, \) is strengthened. This is represented as follows.

\[
\Delta W = \Delta + \exp \left( -\frac{[\Delta(t)]}{\tau_+} \right)
\]

In the case that a spike generates an influence on both directions of increasing and decreasing values, the expression becomes as follows.

\[
\Delta W = A_+ \exp \left( -\frac{[\Delta(t)]}{\tau_+} \right) - A_- \exp \left( -\frac{[\Delta(t)]}{\tau_-} \right)
\]

### 3. Previous work

DSSN model was first implemented on FPGA in [7]. It is based on the circuit architecture shown in Figure 2, and it runs in the pipelined way shown in Figure 3. The FPGA chip used is Virtex 6 XC6VSX315T. The parameter values are shown in Figure 2. Some scalar multiplications are replaced with additions and so the number of multiplications is less than what are shown in the expression 1–4.

In this implementation, according to [7] Accumulator Unit can generate one output every 4 cycles. As shown in Figure 3, DSSN Unit and Silicon Synapse Unit actually operate only 3% of the entire computation time. Nevertheless, Accumulator Unit and other units must exist for parallel computations, as Accumulator Unit generates outputs only at specific small numbers of cycles.

### 4. Our implementation

In this section we introduce our implementation of the DSSN model and its networks. First we show the FPGA system and associated high level synthesis tool that we use for our implementation.

#### 4.1. Target FPGA system

The FPGA system is connected to a host Xeon-based PC through PCI express as shown in Figure 4. The FPGA system is connected to a host Xeon-based PC through PCI express as shown in Figure 4.
board has Virtex 6 XC6VSX475T and 24GB of DRAM memories. The host PC downloads the FPGA configurations and then host PC and FPGA can run at the same time with possible communication. In our implementation, after the DSN models and networks are downloaded into the FPGA, spiking neural networks are simulated on the FPGA board only. After finishing all simulations, the results are transferred to the host PC.

We use a high level synthesis tool, Maxcompiler [8] from Maxeler. It can generate highly pipelined VHDL codes from data flow graphs represented in Java syntax. The VHDL codes are further compiled by Xilinx tools. Maxcompiler does not accept any conditional statements. The inputs to the compiler must be purely data flow graphs for highly pipelining. So, the conditional statements shown in the expressions (1–2) are manually converted into the ones which always compute both cases and selecting the right ones through multiplexers. Maxcompiler can accept the target clock speed. The numbers of pipeline stages can easily exceed 1,000 in typical compilations including in this implementation.

4.2. Overall architecture

We are following the previous implementation [7], but improved the performance by 8 times through more intensive pipelining as well as the use of larger FPGA chips. The overall data flow graph for the computation is shown in Figure 5. The inputs to the circuit are the initial values of $I_s$, $W$ and several control signals, and the outputs are updated $I_s$.

Simulations are performed based on the expressions 1–5 for 64 steps with a fixed value of $W$, and then learnings are performed for 2 steps. This process is repeated until the value of $W$ does not change for some time. When there are N neurons, 1 step is computed with $N+12$ clock cycles.

Figure 6 shows the case of $N=256$.

The entire circuit is decomposed into the weighted sum unit, which computes the expression 5, the difference computation unit, which computes the expressions 1–4, the learning unit, and the storage unit. These are different from Figure 2.

The weighted sum unit receives 2$N$ of $W_{ij}$ as inputs, and so for speedy computations $W_{ij}$ are stored in the block RAM in the FPGA chip.

The difference computation unit operates on the variables, $v, n, q, I_s$, which are also stored in the block RAM. The learning unit and the storage unit are checking when a spike happens and uses it for learning.

4.3. Fast computation of weighted sum

The computation for the expression 5 dominates the total computation time. In the previous implementation [7], Accumulator unit does this computation, but it can generate only one output at every 4 cycles, which is the speed bottleneck.

On the other hand, in our implementation, we introduce the pipeline operations shown in Figure 6 so that more pipeline stages can be introduced as the numbers of neurons to be simulated increases. The computation of $I_{stim}$ takes $N+1$ cycles, and then $I_{s1}$ is computed in 11 cycles. This is pipelined with respect to $i$. The computations in terms of dataflow are aligned for easier layout inside the FPGA chip. The exception on this is the communication
from the difference computation unit to the storage unit. Although there are a number of wires for this communication, right now there is no significant delay caused from them.

4.4. Bit width

Time interval, $\Delta t$, and bit-widths of variables are important parameters, as they influence the simulation time and accuracy. A software simulator has been implemented in C to determine those values. The results say $\Delta t$ should be 3/8000, and $v, n, q, I_{stim}$ needs 18 bits and $I_s, W$ needs 16 bits, which are used in our FPGA implementation.

4.5. Experimental results

The Java description based on the above dataflow has around 200 logical lines of codes. It has been synthesized to the FPGA netlists by Max Compiler [8], which runs at 200MHz. The synthesized results are shown in Table 1. There can be up to 768 neurons in one FPGA chip.

Compared with the existing implementation [7], the number of cycles is reduced to 1/4 and clock speed becomes double, and so in total around 8 times speed up has been observed. Our implementation is 280 times faster than the actual brain.

5. Concluding remarks

We have presented our single FPGA chip implementation of DSSN network models which runs at 200MHz. When there are $N$ neurons and we prepare $p$ computing units whose details are shown above, the computation complexity is proportional to $N^2$ and so the computation time becomes $N^2/p$. With the largest Virtex6 chip from Xilinx, we can implement 768 computing units in a single chip.

We are working on implementations with multiple FPGA chips by decomposing the dataflow graph. With 4 FPGA boards which are interconnected by a ring network, we should be able to simulate around 3,000 neurons at almost the same speed as the case of single FPGA chip.

References


External Input-Facilitated Onset of Chaos in Recurrent Neural Networks

Muyuan XU1, Yuichi KATORI2,3, Takashi KOHNO3, Kazuyuki AIHARA3

1Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan,
2The School of Systems Information Science, Future University Hakodate,
3Institute of Industrial Science, The University of Tokyo, Tokyo, Japan
Email: muyuan@sat.t.u-tokyo.ac.jp

Abstract—We investigate the transition from a fixed point state to chaos in a recurrent neural network. We focus on the intensity of external input to the network and show that the transition occurs when the external input increases. On the other hand, the standard deviation of the external input seems to have little impact on the transition.

1. Introduction

Neural networks with random connections have been receiving increasing attention. In the framework known as “reservoir computing” (Jaeger, 2001; Maass et al., 2002), such networks are found computationally powerful due to their nonlinear nature, where the dynamics is chaotic. However, the dynamics should not deviate far from an ordered one, since neural networks are required to hold useful information temporarily. The transition between ordered and chaotic states in recurrent neural networks is therefore crucial for flexible real-time computation.

An interesting work by Rajan et al. (2010) showed that the ongoing chaos can be suppressed by periodic external input with large magnitude. Recently, Kadmon and Sompolinsky (2015) developed systematically the mean field theory for the transition between a fixed point and chaotic fluctuation in randomly connected networks which consist of multiple subnetworks. Their results revealed the critical role of the synaptic gain and the shape of input-output transfer function in the transition of the dynamics to chaos. For large external input, the transfer function they used also predicts a suppression of chaos due to saturation.

However, in physiological experiments (e.g. Stokes et al., 2013), there is often an increase in mean firing rates after the onset of a stimulus, accompanied by the change in neural variability (Churchland et al., 2010). This could be a sign of change in dynamics, induced by enhanced external input. It therefore remains as a question whether this type of transition is allowed in recurrent neural networks.

In this paper, we consider another often-used transfer function, and show that the increasing in external input could also facilitate the onset of chaos.

2. Network Model

We consider a local network, which consists of $N_E$ excitatory and $N_I$ inhibitory recurrently connected neurons. In addition, the network receives external inputs from another $N_R$ excitatory neurons. Following the idea of balanced network (Vreeswijk and Sompolinsky, 1996), we assume that each neuron receives $C_E = 200$ excitatory and $C_I = 400$ inhibitory synapses on average, from local recurrent connections, and further $C_R = 200$ excitatory synapses from a remote region. Therefore, the number of synapses on each neuron is large, but sparse compared to the total number of neurons in the network. It is known that in a spiking model with such assumption, the excitatory and inhibitory inputs to individual neurons dynamically balance each other (Vreeswijk and Sompolinsky, 1996), rendering the spikes to be determined by stochastic fluctuations. Here we use a rate-based model, where the local dynamics of the $i^{th}$ neuron is given by

$$\begin{align*}
\frac{dx_i^k}{dt} &= -\frac{x_i^k}{\tau_k} + \sum_{j=1}^{N_E} f_x^{EE} r_j^E - \sum_{j=1}^{N_I} f_x^{EI} r_j^I + \sum_{j=1}^{N_R} f_x^{ER} r_j^R, \\
r_i^k &= f(x_i^k), k \in \{E,I\},
\end{align*}$$

where $x_i^k$ is the activation of neuron and $r_i^k$ is the corresponding firing rate. The synaptic strengths $f_x^{EE}, f_x^{EI}$, and $f_x^{ER}$ are either zero, which corresponds to non-connected neurons, or drawn independently from certain distributions, with mean and standard deviation (SD) $\mu_x$ and $\sigma_x$, $k \in \{E,I\}, l \in \{E,I,R\}$. We adopt the often-used transfer function $f(x) = \frac{1}{2}(1 + \tanh(x/\theta))$, where $\theta$ indicates the nonlinearity. We make further simplifications by assuming that $\tau_E = \tau_I = \tau = 10$, $\mu^{EE} = \mu^{EI} = \mu^{ER} = \mu^{IE} = 0.2, \mu^{II} = 5 \cdot 0.2, \sigma^{EI} = \mu^{EI}$ (Amit and Brunel, 1997), and $T = 10$. The external inputs are assumed to be constant over time.

3. Methods and Results

Two different states are found in the dynamics (Figure 1). When the intensity of external input is weak, a stable non-zero fixed point exists in the firing rate space, where almost any initial distribution of firing rates will be attracted into this state. The configuration of firing rates in this state is determined by the specific configuration of synapses, and firing rates in the external input. Note that although this configuration is random due to the randomness in synapses and external input, its statistical properties are rather simple, and can be obtained self-consistently. When the intensity of the external input exceeds some critical value, however, the...
fixed point is destabilized and the network state exhibits

deterministic chaos, a common scenario reported in
balanced networks.

Figure 1. Two states in the dynamics. Top: the mean 
and distribution of firing rates in the fixed point state in a 
network with $N_E = 2000, N_I = 400, \mu_R = 0.02$, and $\sigma_R = 0.02$. The red curve indicates the distribution solved from 
the self-consistent equations. Bottom: the mean firing rate 
in a network with $\mu_R = 0.1, \sigma_R = 0.02$, and the logarithm 
of the power versus frequency for the chaotic state.

We first derive the self-consistent equations for the fixed 
point state by setting $\frac{dx^k}{dt} = 0$ for each neuron. Because the 
number of synapses on each neuron is large, the 
contribution of firing rates of individual neurons is small.

Thus the state of each neuron is approximately independent 
with each other, and the summation terms on the r.h.s. of 
(1) can be taken as Gaussian noise. In the limit of large 
networks, the dynamics of each neuron is therefore given as follows:

$$\frac{dx}{dt} = -\frac{x}{\tau} + \eta,$$  \hspace{1cm} (3)

where $x$, driven by a Gaussian term $\eta$, is also Gaussian. 
This observation allows us to describe the state simply with 
the mean $\mu_x$ and SD $\sigma_x$ of $x$. The probability density of the 
firing rate is therefore

$$P(r) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left[ -\frac{(f^{-1}(r) - \mu_x)^2}{2\sigma_x^2} \right] \cdot [f^{-1}(r)]'. \hspace{1cm} (4)$$

Under the assumption of the fixed point state, $\mu_x$ and $\sigma_x$ 
can be obtained as

$$\begin{align}
\mu_x &= \tau f'[C_E + g(C_I + \mu_R) + \mu_R], \\
\sigma_x^2 &= \tau^2 f'[C_E + g^2(C_I + \mu_R)](2\sigma_R^2 + \mu_R^2) \\
&+ \tau^2 f'[C_E(2\sigma_R^2 + \mu_R^2)]. \hspace{1cm} (5) \hspace{1cm} (6)
\end{align}$$

where $\mu_r$ and $\sigma_r$ are the mean and SD of the firing rate, 
and the correlation between synaptic connection $f^k_j$ and 
firing rate $r^k$ vanishes again due to the large number of 
synapses.

Substituting (5), (6) into (4) and integrating for the first 
two moments of the distribution of firing rates provides 
following self-consistent equations:

$$\begin{cases}
\mu_r = \int_0^1 r P(r, \mu_r, \sigma_r) \, dr, \\
\mu_r^2 + \sigma_r^2 = \int_0^1 r^2 P(r, \mu_r, \sigma_r) \, dr.
\end{cases} \hspace{1cm} (7) \hspace{1cm} (8)$$

In the fixed point state, the mean and SD of the firing 
rate distribution can be solved from these equations by 
numerical methods.

However, the chaotic state is much more complicated 
than what can be described by a low-dimensional system. 
We therefore wish to understand this transition to chaos 
from below, by continuously changing a bifurcation parameter 
in the fixed point state. To this end, we solve the 
self-consistent equations (7) and (8) at certain parameter 
values, and then use a continuation method to investigate 
the changes in solutions w.r.t. the intensity of the external 
input. The result is shown in Figure 2. Solutions for either 
the mean or SD of the firing rate distribution exist for finite 
$\mu_R$, indicating a critical value at which the transition occurs.

Figure 2. Dependency of mean and SD of the firing rate 
on mean external input. The solid line and dashed line 
indicate stable and unstable fixed points, respectively. The 
red points are simulation results in a network with $N_E = 
2000, N_I = 400$, and $\sigma_R = 0$. The green, black and 
magenta points denote the solutions used to calculate the 
spectrum in Figure 3. Above the critical value, no fixed 
point exists in the dynamics. The outlier red point at $\mu_R = 
0.05$ is a result due to the finite network used in the 
simulation.

The stability of fixed point state is determined by the 
Jacobian matrix of (1):

$$\mathbf{J} = \frac{1}{\tau} \mathbf{J} f' = \begin{pmatrix}
I_{N_E} & \frac{1}{\tau} \mathbf{I} \\
\mathbf{0} & \mathbf{0}
\end{pmatrix}$$

where the first $N_E$ columns in $J$ correspond to synapses 
from excitatory neurons, the remaining $N_I$ columns 
correspond to synapses from inhibitory neurons, and $\{x_i\}$
arranged accordingly. The fixed point is stable when all eigenvalues of this matrix have negative real parts. Note that $J$ is a random matrix whose non-zero entries are drawn from two different distributions. Here we simply use numerical results to show the changes in it. In order to do this, we first solve $\mu_x$ and $\sigma_x$ for given parameter values, and then sample a sufficiently large matrix according to these statistics and calculate the eigenvalues. The results are shown in Figure 3. Due to the correlation in the entries, the spectrum appears as an ellipse. As the solution of the fixed point moves along the blue curve in Figure 2, both the SD of the firing rate and the radius of the spectrum of the Jacobian matrix increase gradually. The bifurcation emerges when one of the eigenvalues crosses the imaginary axis. After that, the starting fluctuations further add to the SD of the firing rate, resulting in avalanche which forces the system quickly into chaotic dynamics.

**Figure 3. Changes in the spectrum of the Jacobian matrix.** From left to right, calculated at the green ($\mu_R = 0.0298, \mu_r = 0.0099, \sigma_r = 0.0057$), the black ($\mu_R = 0.0467, \mu_r = 0.0120, \sigma_r = 0.0142$) and the magenta ($\mu_R = 0.0220, \mu_r = 0.0108, \sigma_r = 0.0302$) points in Figure 2, with $N_E = 10000$, and $N_J = 2000$.

Finally, we show that the SD of the external input has relatively small impact on this transition. We generate a network randomly with $N_E = 2000$, and $N_J = 400$, and then run the simulation in this network with external inputs of different means and SDs. The result is shown in Figure 4. Each square denotes the resulting variance in the state, with corresponding combination of $\mu_R$ and $\sigma_R$ and random initial conditions. Therefore the zero variance indicates the fixed point state, and the transition can be found only for large $\mu_R$.

**Figure 4. The SD of the external input has minor effect on the transition.** The variance of the resulting state is shown by color.

4. Discussion

In this paper, we studied the transition from the fixed point state to chaos in a recurrent neural network, which is facilitated by an increasing external input. In the fixed point state, we found that the state can be characterized by two statistics that can be solved self-consistently. We showed how the transition occurs numerically, by continuously changing the bifurcation parameter. We also found that the SD of the external input seems not to largely influence the transition.

These results have some interesting implications. First of all, since the transition is independent of the SD of the external input, the degree of freedom in the configuration of the external input could be exploited for information coding. For example, an enhanced external input at the onset of stimulus sets chaos onset, encouraging the trajectory in network dynamics to explore misaligned dimensions, which has been shown crucial in many cognitive tasks (e.g. Raposo et al., 2014; Kaufman et al., 2014). Meanwhile, the information about the initial state is preserved through the specific configuration in the external input, thus the trajectory could either recover the initial state, or converge to some state corresponding to integrated information after the task. On the other hand, as indicated by the outlier red points in Figure 2, the transition to chaos is postponed in a finite network. This is due to the random fluctuation in the network structure, and is sometimes referred to as the quenched noise. This randomness implies that each realistic neural network may have a different coding scheme.

To realize certain cognitive functions, cortical networks are required quite often to dynamically switch between different types of dynamics: a noise-resistant one for information holding and a perturbation-sensitive one for information manipulation. The transition we studied here might therefore provide some insight on how this could be achieved in a recurrent neural network.

Acknowledgement

This work was partially funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan), and JSPS KAKENHI Grant Number 26280093 and 16K00246.

References

Control of Avoidance for Chaos by using Downhill Simplex Method

Kohsuke Yamato†, Akihiro Kawabata†, Mio Kobayashi†, and Tetsuya Yoshinaga‡

†Department of Creative Technology Engineering, National Institute of Technology, Anan College
265 Aoki Minobayashi, Anan 774–0017, Japan
Email: 1122380@st.anan-nct.ac.jp, {kawabata, kobayashi}@anan-nct.ac.jp
‡Institute of Biomedical Sciences, Tokushima University,
3-18-15 Kuramoto, Tokushima 770–8509, Japan
Email: yosinaga@medsci.tokushima-u.ac.jp

Abstract—In this study, simple nonlinear dynamical systems are made robust against undesirable bifurcation and unstable states using a downhill simplex method, which is widely used in the optimization problem. The proposed method is based on the concept of robust bifurcation analysis; namely, a stability index defined for each parameter value is updated so as to optimize the index. The optimization method is verified through numerical experiments. The presented results can be generalized for driving a system to avoid chaos.

1. Introduction

In a previous study, robust bifurcation analysis was proposed for automatically determining system parameters [1]. The method used the maximum absolute number of eigenvalues of the linearized dynamics at the considered fixed point as a stability index. By minimizing the stability index, we could determine the optimum system parameters for system robustness. The robust bifurcation analysis is the method that uses the qualitative bifurcation theory based on the dynamical systems theory and the optimization of a performance index in the control theory. However, the stability index is not differentiable with respect to the parameter. In order to solve this problem, the matrix inequality method [2, 3] was proposed, which was based on the control theory. In this direct approach, optimization is achieved with the nonlinear matrix inequality constraint. The objective function and the constraint condition should be smooth with respect to system parameters because the optimization requires the steepest descent direction or the Newton direction of the objective function.

In the present study, we adopt the downhill simplex method [4] to solve the optimization problem with the stability index based on the local expansion rate. This method offers solutions for nonlinear optimization problems, and it does not require differentiability for objective functions because parameters are updated by using objective functions geometrically. Here, we show the results of applying this the method to Hénon map and Kawakami maps, which are known as two-dimensional discrete-time dynamical systems. Furthermore, we demonstrate that we can avoid the chaos observed in these systems and find the parameters for the systems having high stabilities.

2. Stability optimization problem

Now we consider the discrete-time dynamical system described as follows:

\[ f : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n \]

\[ (x(k), \lambda) \mapsto x(k+1) = f(x(k), \lambda), \]  \hspace{1cm} (1)

where \( \lambda \in \mathbb{R}^m \) is a set of system parameters, and \( x(k) \in \mathbb{R}^n \) denotes internal state variables. \( k \) denotes a discrete-time step. The fixed point \( x^* \) of map \( f \) satisfies \( x^* = f(x^*, \lambda) \). The Jacobian matrix \( Df(x, \lambda) \) of the map \( f \) with respect to the fixed point \( x^* \) is defined by

\[ Df(x, \lambda) := \left. \frac{\partial}{\partial x} f(x, \lambda) \right|_{x=x^*}. \]  \hspace{1cm} (2)

As for the stability of non-periodic points \( x_0 \) of the map \( f \), we regard \( x_0 \) as an \( N \)-periodic point, and we use the finite-time Lyapunov exponent, or the local expansion rate, defined by

\[ \gamma(N, x_0, \lambda) := \frac{1}{N} \log \| Df^N(x_0, \lambda) \|, \]  \hspace{1cm} (3)

for the minimization problem described by

\[ \min_{x_0, \lambda} \gamma(N, x_0, \lambda), \]  \hspace{1cm} (4)

where \( Df^N(x_0, \lambda) \) is the Jacobian matrix of the \( N \)-times map of the map \( f(x_0, \lambda) \).

3. Optimization with downhill simplex method

In this study, we use the downhill simplex method to solve optimization problems. The method uses a simple algorithm and does not require a derivative function to optimize the parameters. Therefore, we can use this method for the optimization of non-differentiable objective functions. The algorithm is concretely described as follows: First, consider an \( (m+1) \)-polyhedron, whose vertices are corresponding to the parameters in the \( m \)-dimensional parametric space. Then, update those vertices iteratively to
minimize the objective function and converge the objective function to the appropriate value. In this session, the algorithm of the downhill simplex method with \( m = 2 \) is summarized.

3.1. Algorithm of downhill simplex method

Let \( a \) and \( b \) stand for the search parameters with \( m = 2 \). In Fig. 1(A), \( a_j \), \( j = 1, 2, \) and 3 denotes the search points, and \( \lambda_i \) stands for a reflection point of one of these search points on the \((a, b)\)-parameter plane. \( \gamma_j \), \( j = 1, 2, \) and 3, and \( \gamma_r \) are the stability indexes, i.e., the local expansion rates at \( a_j \), \( j = 1, 2, 3 \) and \( \lambda_i \), respectively. The reflection point \( \lambda_i \) is obtained under the point reflection, in the midpoint between \( \lambda_1 \) and \( \lambda_2 \), when \( \gamma_1 < \gamma_2 < \gamma_3 \) is satisfied. Note that a larger stability index corresponds to a smaller local expansion rate because they have opposite directions.

Algorithm 1 shows the algorithm of the downhill simplex method used in this study. \( \theta \) indicates the threshold for the stability indexes, and the function swap returns the parameters sorted by the descending order in the stability indexes, which is in the ascending order with respect to \( \gamma_i \).

3.2. Application to discrete-time dynamical systems

The method was applied to 2-dimensional discrete-time dynamical systems, Hénon map and Kawakami map. The dynamics of Hénon map is described by

\[
\begin{pmatrix}
x(k+1) \\
y(k+1)
\end{pmatrix} = \begin{pmatrix} 1 + y(k) - ax(k)^2 \\ bx(k) \end{pmatrix},
\]

and Kawakami map is described by

\[
\begin{pmatrix}
x(k+1) \\
y(k+1)
\end{pmatrix} = \begin{pmatrix} ax(k) + y(k) \\ x(k)^2 + b \end{pmatrix},
\]

where \( a \) and \( b \) are system parameters, and \( x \) and \( y \) indicate internal state variables. First, on the \((a, b)\)-plane, we suppose that a grid spacing of 0.001, let the \( \lambda_j^0 \) at \( i \)th grid point be a set of initial parameters. Then, let the \( \lambda_j^0 \), \( j = 1, 2, \) and 3 at vertices of an equilateral triangle with its gravity at \( \lambda_0 \) be a set of initial search points. Periodic points and non-periodic points observed at the search points are assumed to be high order periodic points, and the parameters are updated to the direction minimizing the local expansion rate based on the downhill simplex method.

4. Simulation results

In the bifurcation diagrams shown in this section, \( G^p \), \( P^p \), and \( NS^p \) indicate the tangent bifurcation, period-doubling bifurcation, and Neimark-Sacker bifurcation of the periodic points, respectively. Figures 2(a) and (b) show overlapped images of the local expansion rate for attractors and
the bifurcation diagrams of periodic points of the Hénon map and Kawakami maps. The colored contour plots present the values of the local expansion rate, as indicated by the color bar. Cold color indicates a small local expansion rate and then high stability.

Figure 3 shows a typical simulation result with the bifurcation diagram of the periodic points for the Hénon map. The parameters $\lambda^i$, obtained with the updates for $\gamma^i$ less than -0.2, are indicated by the small blue dots. The typical trajectories are presented by the red solid lines with arrows. The ends of the line correspond to the initial parameter $\lambda^0$ and the final parameter $\lambda^k$. The updates are made in the direction of the arrow. Because the local expansion rate of the bifurcation or chaotic behavior is large and the stability is low, the search points move from those parameter regions to these directions to minimize the local expansion rates. Multiple dots remaining in the shaded region in Fig. 3 exist in the high-stability parameter region, and their local expansion rates are less than -0.2. Note that the search points are not reached by high-stability parameters.

Figures 4(a)–(c) show the simulation results with $\theta = -0.1, -0.3, \text{ and } -0.4$ for the Hénon map, and Figs. 4(d)–(f) shows those with $\theta = 0, -0.02, \text{ and } -0.1$ for the Kawakami map. The red colored small dots indicate the arrival points at parameter $\lambda^k$. The fact that the small blue or red dots in Figs. 3 and 4 are mainly distributed in the region with the negative local expansion rate in Fig. 2 suggests that our method successfully operates the system to avoid chaos. In addition, choosing the threshold $\theta$, we could control the stabilities of the systems to avoid chaos and low-stability conditions.

5. Conclusion

In this study, robustification of a nonlinear dynamical system is considered, and the downhill simplex method is applied to solve the optimization problems in robust bifurcation analysis for dynamical systems based on the Hénon map and Kawakami maps. The advantage of this method is that it does not require differentiability of the objective functions. The method is shown to be efficient, and it can be generalized for avoiding chaos.

References


Figure 4: Chaos avoidance with downhill simplex method in the case of (a)–(c): Hénon map and (d)–(f): Kawakami Map.


Experimental verification of amplitude death in a pair of double-scroll circuits coupled by a one-way partial time-varying delay connection

Shinnosuke Masamura†, Yoshiki Sugitani‡, Keiji Konishi†, and Naoyuki Hara†

†Department of Electrical and Information Systems, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531 Japan
‡Department of Electrical and Electronic Engineering, Ibaraki University
4–12–1 Nakanarusawa, Hitachi, Ibaraki 316-8511 Japan
Email: yoshiki.sugitani.0301@vc.ibaraki.ac.jp

Abstract—This report experimentally investigates amplitude death in a pair of oscillators coupled by a one-way partial time-varying delay connection: the connection delay in one direction is varied, but the connection delay in the other direction is constant. Our circuit experiments show that the one-way partial time-varying delay connection can induce amplitude death for long connection delay. The analytical results agree with our experimental results.

1. Introduction

The dynamics of various physical, chemical, and engineering systems can be mathematically modeled by coupled oscillators [1]. In coupled oscillators, we observe various nonlinear phenomena such as synchronization, spatiotemporal chaos, chimera state, and so on. One of such phenomena is amplitude death [2], which is a stabilization of homogeneous steady state in diffusively-coupled oscillators. It is analytically shown that this phenomenon never occurs in coupled identical oscillators [3]. However, if there exists connection delay between oscillators, then amplitude death can occur even in coupled identical oscillators [4]. Amplitude death induced by the connection delay has been actively investigated in nonlinear science [2].

Amplitude death has been expected to suppress undesired oscillations in engineering systems such as coupled laser systems [5], dc micro grid [6], and coupled thermoacoustic oscillators [7]. This is because the usage of amplitude death does not need feedback controllers for stabilization. However, for implementation in real systems, amplitude death has one critical problem: if the connection delay is relatively long due to a practical constraint, then amplitude death cannot be induced [4]. In order to overcome this problem, the following three connections have been proposed: a distributed delay connection [8], a multiple delay connection [9], and a time-varying delay connection [10].

The time-varying delay connection would be easier to be implemented and would not cost compared with the other two connections [10]. The time-varying delay connection has been implemented in electronic circuits [11], and the topology and delay independent design procedure of connection parameters has been proposed [12]. However, all the connection delays have to be varied with high frequency. Thus, this connection is difficult to be implemented for large networks with a huge number of oscillators.

In order to defeat this difficulty, we first proposed a one-way partial time-varying delay connection for a pair of oscillators, in which the connection delay in a direction is varied, but that in the other direction is constant [13]. In addition, we proposed a two-way partial time-varying delay connection for networks, in which connection delays between some oscillators are varied, but the others are constant [12]. These partial time-varying delay connections are obviously easier to be implemented in large networks; however, to the best of our knowledge, there have been few reports about experimental investigations on amplitude death induced by the partial time-varying delay connections.

The present report experimentally investigates amplitude death in a pair of oscillators coupled by the one-way partial time-varying delay connection. The well-known double scroll circuit is employed as the oscillator, and the connection delay is mainly implemented by peripheral interface controllers (PICs) and DA converters. Our experiments show that the one-way partial time-varying delay connection can induce amplitude death even for long connection delay. The analytical results agree with our experimental results.

Figure 1: A pair of oscillators coupled by a one-way partial time-varying delay connection.
2. One-way partial time-varying delay connection

Let us consider a pair of \( m \)-dimensional oscillators (see Fig. 1),

\[
\begin{align*}
\dot{x}^{(1,2)} &= F(x^{(1,2)}) + bu^{(1,2)}, \\
\dot{y}^{(1,2)} &= cx^{(1,2)},
\end{align*}
\]

where \( x^{(1,2)} \in \mathbb{R}^m, y^{(1,2)} \in \mathbb{R}^n, \) and \( u^{(1,2)} \in \mathbb{R} \) are respectively the state variables, the output signals, and the input signals of 1-st and 2-nd oscillators. \( b \in \mathbb{R}^m \) and \( c \in \mathbb{R}^{1 \times mn} \) denote the input and output vectors, respectively. We assume that each oscillator has at least one unstable fixed point \( x^* : F(x^*) = 0. \) The input signals \( u^{(1,2)} \) are given by

\[
u^{(1)} = k \left[ y^{(2)}_{\tau(t)} - y^{(1)} \right], \quad \nu^{(2)} = k \left[ y^{(1)}_{\tau_0} - y^{(2)} \right],
\]

where \( y^{(2)}_{\tau(t)} := y^{(2)}(t-\tau(t)) \) and \( y^{(1)}_{\tau_0} := y^{(1)}(t-\tau_0) \) are delayed output signals. \( k > 0 \) denotes the coupling strength. \( \tau_0 > 0 \) is the constant delay, and \( \tau(t) > 0 \) denotes the periodically time-varying delay (see Fig. 1) around the nominal delay \( \tau_0 \) with the amplitude \( \delta \in [0, \tau_0) \).

\[
\tau(t) := \tau_0 + \delta f(\Omega t),
\]

where \( \Omega > 0 \) is the frequency of a periodic sawtooth function \( f(x) \),

\[
f(x) := \begin{cases} 
\frac{2x}{\pi} - 1 - 4n & \text{if } x \in [2n\pi, (2n+1)\pi) \\
\frac{2x}{\pi} + 3 + 4n & \text{if } x \in [(2n+1)\pi, 2(n+1)\pi)
\end{cases}, 
\]

\[n = 0, 1, 2, \ldots.
\]

We will consider the local stability of a homogeneous steady state in a pair of oscillators (1), (2),

\[
\begin{bmatrix} x^{(1)T} \ x^{(2)T} \end{bmatrix}T = \begin{bmatrix} x^{T} \ x^{2T} \end{bmatrix}T.
\]

Substituting the perturbation \( \Delta x^{(1,2)} := x^{(1,2)} - x^* \) into Eqs. (1), (2), it yields the dynamics around steady state (4),

\[
X = \{ I_2 \otimes (A - kbc) \} X + B_1 X + B_2 X_{\tau(t)},
\]

where \( A := [-\partial F(x)/\partial x]_{x=x^*} \) is the Jacobian matrix, and

\[
X := \begin{bmatrix} \Delta x^{(1)} \\ \Delta x^{(2)} \end{bmatrix}, \quad B_1 := \begin{bmatrix} 0 & 0 \\ kbc & 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} 0 & kbc \\ 0 & 0 \end{bmatrix}.
\]

For sufficiently large \( \Omega \), the stability of linear time-varying system (5) is guaranteed if linear time-invariant system

\[
\dot{X} = \{ I_2 \otimes (A - kbc) \} X + B_1 X + B_2 \frac{\tau_0}{2\delta} \int_{t-\tau_0-\delta}^{t-\tau_0+\delta} X(s) \, ds,
\]

is stable [14]. Thus, we will focus on the stability of system (6) instead of system (5). The characteristic equation of system (6) is given by,

\[
G(s) := \det \left[ sI_{2m} - \{ I_2 \otimes (A - kbc) \} - B_1 e^{-\tau_0} - B_2 e^{-\tau_0} H(s) \right] = 0,
\]

where

\[
H(x) := \frac{\sin x}{x} \quad \text{if } x \neq 0, \\
H(x) := 1 \quad \text{if } x = 0.
\]

Therefore, the local stability of steady state (4) is governed by the roots of Eq. (7).

For checking the stability of characteristic Eq. (7), we derive the marginal stability curves on the connection parameter \( k, \tau_0 \) space. Substituting \( s = i\lambda (\lambda \in \mathbb{R}) \) into Eq. (7), we obtain,

\[
G(i\lambda) = G_+(\lambda)G_-(\lambda) = 0,
\]

where

\[
G_\pm(\lambda) := \det \left[ i\lambda I_m - A + kbc \pm ke^{-\lambda t_0} \Phi(\lambda \delta)bc \right].
\]

Focusing on the stability of the homogeneous steady state; hence, we cannot deal with the global stability.

3. Experimental circuits

In our experiments, we deal with coupled double-scroll circuits [15] (see Fig. 2),

\[
\begin{align*}
\frac{dv_1^{(1,2)}}{dt} &= \frac{1}{R} (v_2^{(1,2)} - v_1^{(1,2)}) - h(v_1), \\
\frac{dv_2^{(1,2)}}{dt} &= \frac{1}{R} (v_1^{(1,2)} - v_2^{(1,2)}) + h(v_2), \\
\frac{dv_1^{(1,2)}}{dt} &= \frac{1}{L} (i_1^{(1,2)} + i_2^{(1,2)}) + \frac{i_u}{L},
\end{align*}
\]

where \( v_1^{(1,2)} [V] \) and \( v_2^{(1,2)} [V] \) denote the voltages of capacitors \( C_1 [F] \) and \( C_2 [F] \), respectively, \( i_1^{(1,2)} [A] \) is the current through inductor \( L [H] \). The current \( h(v_1) [A] \) through the nonlinear resistor is given by

\[
h(v) := m_0 v + \frac{1}{2} (m_1 - m_0) |v + B| + \frac{1}{2} (m_0 - m_1) |v - B|.
\]

The two circuits are coupled through the coupling resistor \( r \). Thus, the coupling signals are given by

\[
i_u^{(1)} = \frac{1}{r} (v_2^{(1,2)} - v_2^{(1)}), \quad i_u^{(2)} = \frac{1}{r} (v_2^{(1)} - v_1^{(2)}).
\]
Figure 2: Experimental circuit diagram.

where $v_1^{(2)} := v_1^{(2)}(t - \tau(t))$ and $v_2^{(1)} := v_2^{(1)}(t - \tau_0)$ are the delayed voltages. These delayed voltages are generated by the delay units in Fig. 2, which are implemented by PICs (PIC18F2550) and DA converters [11].

The non-dimensional form (1), (2) of coupled double scroll circuits (9), (10) is given with

$$\begin{align*}
F(x) &= \begin{bmatrix}
\eta (x_2 - x_1 - g(x_1)) \\
x_1 - x_2 + x_3 \\
-\gamma x_2
\end{bmatrix}, \\
b &= \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \\
c &= \begin{bmatrix} 0 \end{bmatrix}^T,
\end{align*}$$

(11)

where

$$\begin{align*}
x_1 &= \frac{v_1}{B_p}, \\
x_2 &= \frac{v_2}{B_p}, \\
x_3 &= \frac{i_L R}{\gamma R}, \\
k &= \frac{R}{r}, \\
\eta &= \frac{C_2}{C_1}, \\
\gamma &= \frac{R^2 C_2}{L}, \\
\eta &= m_1 R, \\
a &= m_0 R, \\
b &= m_0 R,
\end{align*}$$

and

$$g(x) := bx + \frac{1}{2} (b - a) |x - 1| - |x + 1|.$$  

Note that in Eq. (11), the non-dimensional time $t/(RC)$ is used instead of the real time $t$. The double scroll circuit (11) has the three equilibrium points, $x^*_2 \approx \pm p$ and $x^*_0 := 0$, where $p := (b - a)/(b + 1)$. Here, we focus on the stability of $x^*_2$. The Jacobian matrix around $x^*_2$ is given by

$$A = \begin{bmatrix}
-\eta (b + 1) & \eta & 0 \\
1 & -1 & 1 \\
0 & -\gamma & 0
\end{bmatrix}.$$  

4. Experimental results

The parameters of circuit (9) are fixed at

$$\begin{align*}
C_1 &= 0.1 \times 10^{-6} \text{ F}, \\
C_2 &= 1.0 \times 10^{-6} \text{ F}, \\
L &= 180 \times 10^{-3} \text{ H}, \\
R &= 1, 800 \Omega, \\
B_p &= 1.0 \text{ V}, \\
m_0 &= -0.4 \times 10^{-3}, \\
m_1 &= -0.8 \times 10^{-3}.
\end{align*}$$

(12)

The circuit with parameter (12) shows the well-known double scroll attractor [15]: a resonant frequency of each oscillator at the equilibrium point is approximately 3.45. The frequency of time-varying delay (3) is fixed at a large value $\Omega = 23$.

Figure 3 shows the stability regions (i.e., shaded areas) for (a) $\delta = 0$ (i.e., time-invariant delay connection) and (b) $\delta = 0.35$ (i.e., one-way partial time-varying delay connection) on $(k, \tau_0)$ space. These regions are derived from the marginal stability curves which are solutions of Eq. (8). Comparing the region in Fig. 3(a) with that in Fig. 3(b), we see that the one-way partial time-varying delay connection expands the region substantially. Especially, for $k > 7.1$ in Fig. 3(b), there are no curves; that is, we can use long connection delay $\tau_0$ to induce amplitude death.

The symbol $\bigcirc$ (×) in Fig. 3 denotes the occurrence (non-occurrence) of amplitude death experimentally. It can be confirmed that most of the experimental results (i.e., symbols $\bigcirc$ and ×) agree with our analytical results (i.e., the shaded areas). A few parameter sets of experimental results do not agree with the analytical results because of parameters mismatch between the two oscillators. Figure 4 shows the time-series data of the voltages $v_1^{(1)}$ and $v_2^{(1)}$ at point A: $(k, \tau_0) = (7.80, 1)$ and point B: $(k, \tau_0) = (2.85, 3)$ in Fig. 3(b). The two double-scroll circuits are coupled at $t = 60$ ms; that is, the switch SW in Fig. 2 is turned on. For point A, the voltages converge onto the equilibrium point after coupling. On the other hand, for point B, the voltages continue to oscillate even after coupling.1

5. Conclusion

This report has experimentally investigated amplitude death induced by a one-way partial time-varying delay connection in a pair of double scroll circuits. It has been experimentally verified that the one-way partial time-varying delay connection can induce amplitude death even for long connection delay. Our experimental results agreed well with analytical results.

1The time-series data in Fig.4(b) seems to be a periodic solution. Since our results are based on the local stability of steady state (4), we cannot deal with global behavior, such as the periodic solution.
(a) \( \delta = 0 \) (time-invariant delay connection)

(b) \( \delta = 0.35 \) (one-way partial time-varying delay connection)

Figure 3: Stability regions (i.e., shaded areas) for a pair of double scroll circuits. The symbol \( \bigcirc \) (\( \times \)) denotes the occurrence (non-occurrence) of amplitude death in our experiments (\( \Omega = 23 \)).

Acknowledgement

The present research was partially supported by JSPS KAKENHI (26289131).

References

An Implementation of Fault-Tolerant FSSP Algorithms

Naoki KAMIKAWA† and Hiroshi UMEO†
† Univ. of Osaka Electro-Communication, Neyagawa-shi, Hatsu-cho, 18-8, Osaka, 572-8530, Japan

e-mail: naoki@osakac.ac.jp, umeo@cyt.osakac.ac.jp

Abstract
We study a classical firing squad synchronization problem (FSSP) on a model of fault-tolerant cellular automata that might have possibly some defective cells. Several fault-tolerant nearly minimum-time FSSP algorithms are implemented on finite-state automata. It is shown that, under some constraints on the distribution of defective cells, any cellular array of length \( n \) with \( p \) defective cell segments can be synchronized in \( 2n - 2 + p \) steps. The FSSP algorithms developed are realized on a finite state automaton with 164 states and 4792 rules.

1. Introduction
Synchronization of large-scale networks is an important and fundamental computing primitive in parallel and distributed systems. The synchronization in ultra-fine grained parallel computational model of cellular automata, known as firing squad synchronization problem (FSSP), has been studied extensively for more than fifty years, and a rich variety of synchronization algorithms has been proposed. In the present paper, we consider the FSSP from a viewpoint of fault tolerance. Reliable and fault-tolerant computation in a large scale of cellular automata is a key issue to be studied so far. Gács [1] constructed reliable cellular automata from unreliable ones that make errors with some constant probability. Fault tolerance in FSSP has been studied by Kutrib and Vollmar [2], Umeo [4] and Yunès [6]. One of the major open questions has been: How many states are required in their realizations on a finite state automaton? No implementations were made in the past.

In this paper, we study the FSSP for 1D cellular automata with specific defective patterns such that there are locally more intact cells than defective ones. It is shown that there exist several time-efficient synchronization algorithms with fault-tolerance and give the first implementation in terms of finite state automaton. Section 2 gives a description on a fault-tolerant model of cellular array. Section 3 presents several time-efficient FSSP algorithms with fault tolerance and their implementations on a finite state automaton with 164 states and 4792 rules.

2. Fault-Tolerant FSSP

2.1. FSSP
Consider a finite 1D cellular array consisting of \( n \) cells. Each cell is an identical (except the border cells) finite-state automaton. The array operates in lock-step mode in a way such that the next state of each cell (except border cells) is determined by both its own present state and the present states of its left and right neighbors. All cells (soldiers), except the left end cell (general), are initially in a quiescent state at time \( t = 0 \). The quiescent state has a property that the next state of a quiescent cell with quiescent neighbors is the quiescent state again. At time \( t = 0 \), the left end cell \( C_1 \) is in the fire-when-ready state, which is the initialization signal for the array. The firing squad synchronization problem is to determine a description (state set and next-state function) for cells that ensures all cells enter the fire state at exactly the same time and for the first time. The set of states and the next-state function must be independent of \( n \).

2.2. Delayed Synchronization
First we introduce a freezing-thawing technique that yields a delayed synchronization.

Lemma 1 Let \( t_1 \), \( t_2 \) and \( \Delta t \) be any integer such that \( 0 \leq t_1 \leq n - 1 \), \( t_1 \leq t_2 \) and \( \Delta t = t_2 - t_1 \). We assume that the right end cell of the array of length \( n \) receives a special signal from outside at time \( t = t_1 \) and \( t_2 \). Then, there exists a CA that can fire at time \( t = 2n - 2 + \Delta t \).

The array operates as follows:

1. Start a minimum-time FSSP algorithm at left end of the array. A freezing signal is given from outside at time \( t = t_1 \) at the right end of the array. The signal is propagated in the left direction at its maximum speed, that is, 1 cell per 1 step, and freezes the configuration progressively. Any cell that receives the freezing signal from its right neighbor has to stop its state-change and transmits the freezing signal to its left neighbor. The frozen cell keeps its state as long as no thawing signal will arrive.

2. A special signal supplied with outside at time \( t = t_2 \) is used as a thawing signal that thaws the frozen configuration. The thawing signal forces the frozen cell to resume its state-change procedures immediately. The signal is also transmitted toward the left end at speed 1/1.

We can freeze the entire configuration during \( \Delta t \) steps and delay the synchronization on the array for \( \Delta t \) steps. We refer the scheme as freezing-thawing technique.
2.3. Cellular Automata with Defective Cells

2.3.1. Intact and Defective Cells

Consider a 1D array of cells, some of which are defective. Each cell has its own self-diagnosis circuit that diagnoses itself before its operation. The diagnosis result is stored as a flag in the special register augmented with each cell. We assume that new defections do not occur during the operational lifetime on any cell. A consecutive defective (intact) cells are referred to as a defective (intact) segment, respectively. Figure 1 illustrates a 1D array with three defective and four intact segments. Any defective and intact cells can detect whether its neighbor cells are defective or not. Without loss of generality, we can restrict our investigation to cellular arrays that have an intact segment at its left and right ends, shown in Fig. 1.

We use the following notations. The array consists of \( p \) defective segments and \((p + 1)\) intact segments, denoted by \( I_1 \) and \( D_1 \), respectively and \( p \) be any positive integer. Let \( n_1 \) and \( m_1 \) be number of cells on the \( i \)th intact and \( j \)th defective segments, where \( i \) and \( j \) be any integer such that \( 1 \leq i \leq p + 1 \) and \( 1 \leq j \leq p \). Let \( n \) be the number of cells of the array such that \( n = (n_1 + m_1) + (n_2 + m_2) + \ldots + (n_p + m_p) + n_{p+1} \). Throughout this paper, we study the synchronization algorithms for arrays such that there exist locally more intact cells than defective ones, i.e., \( n_i \geq m_i \) for any \( i \) such that \( 1 \leq i \leq p \).

2.3.2. Signal Propagation in a Defective Segment

In our model we assume that any cell in defective segment can only transmit the signal to its right or left neighbor depending on the direction in which it comes to the defective segment. The speed of the signal in any defective segment is fixed to 1/1, that is, one cell per one step. In defective segments, both the information carried by the signal and the direction in which the signal is propagated are preserved without any modifications. Thus, we can see that any defective segment has two one-way pipelines that can transfer one state at 1/1 speed in either direction. Note that from a viewpoint of state transition of CAs each cell in a defective segment can change its internal states in a specific manner. The readers find its formal definition in Kutrib and Vollmar [1995] and Yunis [1996]. Figure 2 shows how two signals with different speeds and directions propagate in a defective segment.

2.4. Fault-Tolerant FSSP

The fault-tolerant FSSP for cellular automata with defective cells is to determine a description for cells that ensures all intact cells enter the fire state at exactly the same time and for the first time. It is easily seen that a lower bound of time complexity for synchronizing any array with some defective segments of length \( n \) is \( 2n - 2 \).

3. Fault-Tolerant FSSP Algorithms

In this section we give several versions of fault-tolerant FSSP algorithms. First we consider the case \( p = 1 \) where the array has one defective segment and \( n_1 \geq m_1 \).

Synchronization Algorithm \( A_1 \): In order to synchronize two intact segments simultaneously at time \( t = 2n - 2 \), a minimum-time synchronization algorithm is applied independently to \( I_1 \) and \( I_2 \). Our previous freezing-thawing technique is employed for the synchronization of \( I_1 \), where the configuration on \( I_1 \) is frozen during \( \Delta t \) (given below) steps to keep pace with the synchronization for \( I_2 \). As for the synchronization of \( I_2 \), on the other hand, an minimum-time synchronization algorithm without any delay is used. Figure 3 illustrates the synchronization scheme. At time \( t = 0 \), the general located at left end of the array generates two signals which propagate in the right direction at speed 1/1 and 1/3, respectively. The fast signal continues to move to the right end of the array, through \( I_1, D_1 \) and \( I_2 \). When the signal arrives at each right end of \( I_1, D_1 \) and \( I_2 \), it generates a reflected signal which propagates at 1/1-speed in the left direction. The first reflected signal acts as an initialization for freezing the configuration on \( I_1 \). The second one goes through \( D_1 \) and \( I_1 \) and meets the 1/3-speed signal at time \( t = \lceil 3(m_1 + n_1)/2 \rceil \) on the middle cell \( C_{\lceil (m_1 + n_1)/2 \rceil} \) of \( I_1 \) and \( D_1 \). Note that it belongs to \( I_1 \), since \( n_1 \geq m_1 \). The third one, going through \( I_2 \) and \( D_1 \), arrives at the right end of \( I_2 \) at time \( t = n + m_1 + n_2 - 1 \), and it initiates a thawing operation for \( I_1 \). Thus the configuration on \( I_1 \) is frozen during \( \Delta t = (n + m_1 + n_2 - 1) - (n_1 - 1) = n + m_1 + n_2 - n_1 \) steps.

Based on Lemma 1, the first intact segment \( I_1 \) can be fired at time \( t = 2n_1 - 2 + \Delta t = 2n - 2 \) in minimum-step. On the other hand, the synchronization for \( I_2 \) is initiated at time \( t = 2(n_1 + m_1) \) by the right-going 1/1-speed signal generated by the cell \( C_{\lceil (m_1 + n_1)/2 \rceil} \) at time \( t = \lceil 3(m_1 + n_1)/2 \rceil \). See Fig. 3. The segment fires \( 2(n_2 - 2) \)-step later, that is, at time \( t = 2(m_1 + n_1) + 2(n_2 - 1) = 2n - 2 \). Thus, the entire intact segments can be synchronized at time \( t = 2n - 2 \) in minimum-time. Thus we have:

**Theorem 2** Let \( M \) be any cellular array of length \( n \) with
one defective and two intact segments. Then, $M$ is synchronizeable in $2n - 2$ minimum-time.

![Diagram of synchronization algorithm A1](image)

**Figure 3:** A space-time diagram for minimum-time FSSP algorithm with one defective segment (left) and its implementation (right).

We can generalize the synchronization scheme $A_1$ to arrays with multiple defective segments more than two. Now we consider an array with a fixed number of defective segments. In this case, each cell knows the number $p$ in advance. Precisely, Let $p$ be a fixed positive integer and $M$ be any cellular array of length $n$ with $p$ defective segments, where $n_i \geq m_i$ for any $i$ such that $1 \leq i \leq p$.

**Synchronization Algorithm $A_2$:** Each intact segment $I_i$, $1 \leq i \leq p + 1$, can be synchronized by the similar method employed in the synchronization algorithm $A_1$. As for the first intact segment, the same operations are performed as in $A_1$, with the exception of the thawing signal. At time $t = 0$, the general generates two signals, propagating in the right direction at speed 1/1 and 1/3, respectively. The fast signal continues to move to the right end of the array. When the fast signal arrives at each right end of $I_1$, $D_1$ and $I_{p+1}$, it generates a reflected signal propagating at 1/1-speed in the left direction. The first reflected signal generated by the right end of $I_1$ at time $t_1 = n_1 - 1$ begins to freeze the configuration on $I_1$. The second one goes through $D_1$ and $I_1$ and meets the 1/3-speed signal at time $t = [3(m_1 + n_1)/2]$ on the cell $C_1 = [m_1 + n_1/2]$ that belongs to $I_1$, since $n_1 \geq m_1$. In addition, the cell sends out a 1/1-speed right-going signal that will initiate the next general for $I_2$. The third one, going through $I_{p+1}$, $D_{p+1}$, $I_p$, $D_p$, ..., $D_3$, $I_2$ and $D_1$, arrives at the right end of $I_1$ at time $t_2 = t_1 + 2m_1 + 2 \sum_{j=2}^{p} (n_j + m_j) + 2n_{p+1}$, and it initiates a thawing operation for $I_1$. The configuration on $I_1$ has been frozen during $\Delta t = t_2 - t_1 = 2m_1 + 2 \sum_{j=2}^{p} (n_j + m_j) + 2n_{p+1}$ steps. Thus, based on [Lemma 1], $I_1$ can be fired at time $t = 2m_1 - 2 + \Delta t = 2 \sum_{j=1}^{p} (n_j + m_j) + n_{p+1} - 2 = 2n - 2$.

For any $i$ such that $2 \leq i \leq p$, we have the following observations:

1. A 1/1-speed right-going signal generated by the middle cell of $I_{i-1}$ and $D_{i-1}$ generates a general at the left end of $I_i$ at time $t_i = 2 \sum_{j=1}^{i-1} (n_j + m_j)$. It initiates the synchronization of $I_i$. At time $t = t_i$, the general generates two signals which propagate in the right direction at speed 1/1 and 1/3, respectively. The fast signal continues to move to the right end of the array. When the signal arrives at each right end of $I_i$, $D_i$ and $I_{i+1}$, it generates a reflected signal which propagates at 1/1-speed in the left direction.

2. The first reflected signal generated at the right end of $I_i$ at time $t_{i+1} = t_i + n_i$ begins to freeze the configuration on the segment.

3. The second one goes through $D_i$ and $I_i$ and meets the 1/3-speed signal at time $t = t_i + [3(m_i + n_i)/2]$ on the cell $C_t = [3(m_i + n_i)/2]$ that belongs to $I_i$, since $n_i \geq m_i$. In addition, the cell sends a 1/1-speed right-going signal that will initiate the next general for $I_{i+1}$.

4. The third reflected signal, going through $I_{i+1}$, $D_i$, ..., $D_j$, $I_j$, $j = p, p-1, ..., i+1$, and $D_i$ in order, arrives at the right end of $I_i$ at time $t = t_i + 2m_i + 2 \sum_{j=i+1}^{p} (n_j + m_j) + 2n_{p+1}$, and it initiates a thawing operation for $I_i$. Based on [Lemma 1], the $i$-th intact segment $I_i$ can be fired at time $t = t_i + 2n_i - 2 + \Delta t = 2n - 2$.

It is easily seen that the last intact segment can be fired at time $t = t_{p+1} + 2n_{p+1} - 2 = 2n - 2$. To distinguish $p$ thawing signals at each right end of $I_i$, the rightmost cell of the array assigns an identified code to each signal. Thus, the entire intact segments can be synchronized at time $t = 2n - 2$ in minimum-step. Note that the number of internal states of the cells required for realizing the algorithm $A_2$ depends on $p$. Figure 4 shows the time-space diagram for the case $p = 3$.

![Diagram of synchronization algorithm A2](image)

**Figure 4:** A space-time diagram for minimum-time firing squad synchronization algorithm with three defective segments (left) and its implementation (right).

Thus we have:

**Theorem 3** Let $p$ be a fixed positive integer and $M$ be any cellular array of length $n$ with $p$ defective segments, where $n_i \geq m_i$ for any $i$ such that $1 \leq i \leq p$. Then, $M$ is
synchronizable in $2n - 2$ minimum-time.

Generally speaking, there are at most $O(n)$ intact and defective segments on a cellular array of length $n$. In the case where $p$ is any positive integer, our algorithm $A_2$ cannot assign a unique identification code to each thawing signal within a finite state description. Thus the rightmost end cell in each intact segment cannot distinguish those $p$ left-going return signals that act as a thawing signal for each corresponding intact segment. Even in this case we can get a rather slower algorithm with placing a small restriction on the length of defective and intact segments. Let $p$ be any positive integer and $M$ be any cellular array of length $n$ with $p$ defective segments, where $n_i \geq m_i$ and $n_i + m_i \geq p - i$, for any $i$ such that $1 \leq i \leq p$.

**Synchronization Algorithm $A_1$:** We need a few modifications to the algorithm $A_2$. First two new thawing signals, that is, $a$- and $b$-signals, are introduced in order to thaw the right intact segment. Second, the initiation of the synchronization process is delayed for one step at each intact segment. Precisely, the synchronization for the $i$-th segment is initiated at time $t_i = 2 \sum_{j=1}^{n-1} (n_j + m_j) + (i - 1)$. Whenever the fast signal (the same signal as described in $A_2$) arrives at each right end of intact segment, it splits into two signals. One is the freezing signal and the other is the $a$- and $b$-signals which propagate toward the right end of the array at $1/1$-speed. The $b$-signal stays for one step at the left end of each intact segment that it encounters. Both $a$- and $b$-signals reflect at the right end of the array and proceed to the left direction at $1/speed$. This time the reflected $a$-signal stops for one step at the left end of each defective segment that it encounters. When the following conditions given below are satisfied, two reflected $a$- and $b$-signals meet at the right end of right intact segment just where the original $a$- and $b$-signals have been generated. Now the thawing operation for the configuration of the intact segment is started.

Let $t_i^a$ and $t_i^b$ be time at which the $a$- and $b$-signals emitted by the $i$-th segment hit the right end of the array, respectively. We have:

$$t_i^a = t_i + \sum_{j=1}^{p} (n_j + m_j) + p^+ + 1,$$

$$t_i^b = t_i^a + p + i + 1.$$

The freezing and thawing operations for $I_i$ are started, respectively, at time $t_i = t_i^a + n_i + 1$ and $t_i = t_i + 2m_i + 2 \sum_{j=i+1}^{n} (n_j + m_j) + 2n^+ + p + i + 1$.

The condition: $t_i^{a,b} \geq t_i$ for any $i$ such that $1 \leq i \leq p$ is necessary and sufficient for the configuration on $I_i$ to be thawed by the thawing signal emitted by the $i$-th segment. The condition is satisfied for any such $i$ that $1 \leq i \leq p$, since $t_i^{a,b} = n_i + m_i - p + i \geq 0$.

Thus the configuration on $I_i$ is frozen during $\Delta t = t_i - t_i = 2m_i + 2 \sum_{j=i+1}^{n} (n_j + m_j) + 2n^+ + p + i + 1$ steps. Based on Lemma 1, the $i$-th intact segment $I_i$ can be fired at time $t = t_i + 2n_i - 2 + \Delta t = 2n - 2 + p$. In this way, the entire intact segments can be synchronized at time $t = 2n - 2 + p$. From the assumptions $n_i + m_i \geq p - i$, for any $i$, $1 \leq i \leq p$, it is seen that $p = O(\sqrt{n})$. Thus the time complexity of the algorithm is $2n + O(\sqrt{n})$. Figure 5 shows the synchronization scheme for a cellular array with three defective segments. Thus we have:

**Theorem 4** Let $p$ be any positive integer and $M$ be any cellular array of length $n$ with $p$ defective segments, where $n_i \geq m_i$ and $n_i + m_i \geq p - i$, for any $i$ such that $1 \leq i \leq p$. Then, $M$ is synchronizable in $2n - 2 + p$ steps.

![Figure 5: A space-time diagram for minimum-time firing squad synchronization algorithm with three defective segments (left) and its implementation (right).](image)

4. Conclusions

It has been shown that, under some constraints on the distribution of defective cells, any cellular array of length $n$ with $p$ defective cell segments can be synchronized in $2n - 2 + p$ steps. Several implementations have been given for the fault-tolerant FSSP algorithms together with a realization on a finite state automaton having $164$ states and $4792$ rules.

References


Steady State Analysis of Digital Return Maps and Cellular Automata

Naoki Tada†, Hiroki Yamaoka‡ and Toshimichi Saito†

†EE Dept., Koganei, Tokyo 184-8584, Japan, Email: tsaito@hosei.ac.jp
‡Hitachi Kokusai Electric Inc., Chiyoda-ku, Tokyo 101, Japan

Abstract—This paper considers dynamic of digital return maps, simple digital dynamical systems on a set of points. Depending on parameters and initial states, the map can generate various periodic orbits. In order to analyze the steady state, we present two feature quantities: periodic orbits occupancy rate and dispersion of periods of the orbit. Using the feature quantities, we construct the feature quantities plane that is useful in visualization and classification of the dynamics. Using the feature quantities, a simple class of cellular automata are analyzed.

1. Introduction

A digital return map (Dmap [1] [2]) is a simple digital dynamical system on a set of points. It can be regarded as a digital version of analog one-dimensional maps represented by the logistic map [3] [4]. Since the number of points is finite, the Dmap cannot generate chaos. However, depending on parameters and initial states, the Dmap can generate a variety of periodic orbits (PEO) in the steady state.

The Dmap is related to various digital dynamical systems and their applications including cellular automata and signal/image processing [5]-[8], and dynamic binary neural networks and control of switching circuits [9]. Analysis of the Dmap can contribute not only to basic study of nonlinear dynamics but also to engineering applications. However, the analysis is hard because the Dmap has a large variety and the dynamics is very complicated.

In order to analyze steady states of Dmaps, this paper presents two simple feature quantities. The first quantity is occupancy rate of the PEOs for all the initial points. It can characterize plentifulness of the steady states. The second quantity is dispersion of periods of the PEOs. It can characterize variety of the PEOs. Using the two feature quantities, a feature plane is constructed. The feature plane is useful in visualizing and classification of the dynamics.

As an example of the Dmaps, we consider a Dmap derived from a simple class of elementary cellular automata (ECAs) on 8 dimensional binary spaces. The dynamics of ECA is governed by one rule and can generate a variety of spatiotemporal patterns. The dynamics of the ECAs is integrated into a Dmap on a set of $2^8$ points. Calculated the feature quantities of the ECA, we visualize/classify rich dynamics of the ECA. Note that this is the first publication of the feature plane for steady state analysis and its application to ECAs.

2. Digital return map and simple feature quantities

The Dmap is defined on a set of points to itself and its iteration generates a digital sequence as shown in Fig. 1:

$$\begin{align*}
\theta_{n+1} &= f(\theta_n), \quad \theta_n \in L_N \\
L_N &\equiv \{l_1, \cdots, l_N\}, \quad l_i = i/N, \quad i = 1 \sim N
\end{align*}$$

where $\theta_n$ is a digital state variable on $L_N$ at discrete time $n$ and $L_N$ is a set of $N$ points $l_i$. Since the points are equivalent to binary vectors, we refer to this systems as to be digital.

Since the number of points in the domain $L_N$ is finite, steady state must be a periodic orbit. Here we give basic definitions.

**Definition 1**: A point $\theta_p \in L_N$ is said to be a periodic point with period $p$ if $f^p(\theta_p) = \theta_p$ and $f^q(\theta_p) \neq \theta_p$ for $0 < q < p$ where $f^p(\theta)$ is the $p$-composition of $f$. A PEO with period 1 is referred to as a fixed point. A sequence of the periodic points $(f(\theta_p), \cdots, f^p(\theta_p))$ is said to be a periodic orbit (PEO) with period $p$.

**Definition 2**: A point $\theta_e$ is said to be an eventually periodic point (EPP) if $\theta_e$ is not a periodic point and there exists some positive integer $k$ such that $f^k(\theta_e)$ is a periodic point.

The Dmap for $N = 16$ in Fig. 1 has one fixed point and one PEO with period 3. The other 12 points are EPPs.

In order to consider the steady state dynamics, we introduce two simple feature quantities. First, let $N_p$ be the number of periodic points on $N$ points. The first quantity is

![Figure 1: A digital return map (Dmap) and distribution of periods. The blue point is a fixed point and $P_1 = 1/4$. The red orbit is a periodic orbit (PEO) with period 3 and $P_2 = 3/4$. Black points are eventually periodic points (EPPs).](image-url)
the PEOs occupancy rate over the domain of the Dmap

\[ \alpha = \frac{N_p}{N}, \quad \frac{1}{N} \leq \alpha \leq 1 \]  

(2)

This quantity characterize plentifulness of steady states.

In order to define the second feature quantity, we define several notations. Let a Dmap have \( N_p \) pieces of periodic points. Let \( N_e \) be the number of PEOs and let \( p_i \) be the period of the \( i \)-th PEO, \( i = 1 \sim N_e \). Let \( P_i = p_i/N_p \) where \( \sum_{i=1}^{N_p} P_i = 1 \). \( \{P_i\} \) is referred to as a period distribution. The second quantity \( \gamma \) is dispersion of periods of the PEOs.

\[ \gamma = \sum_{i=1}^{N_p} P_i^2 \alpha, \quad \frac{1}{N} \leq \gamma \leq 1 \]  

(3)

It can characterize variety of PEOs. If a Dmap has one PEO then \( \gamma \) takes the maximum value \( \gamma = 1 \). If a Dmap has \( N \) fixed points then \( \gamma \) takes the minimum value \( \gamma = 1/N \).

In the Dmap in Fig. 1, one fixed point and one PEO with period 3 correspond to \( P_1 = 1/4 \) and \( P_2 = 3/4 \), respectively. Hence this Dmap is characterized by \( \gamma = 10/16 \) and \( \alpha = 4/16 \).

In order to consider the dynamics, the \( \alpha \) versus \( \gamma \) feature plane is constructed as shown in Fig. 2. On the \( \alpha-\gamma \) feature plane, we define three criterial objects for consideration of the dynamics. The first criteria is defined for a Dmap in Fig. 2 (a): if a Dmap has only one PEO then \((\alpha, \gamma)\) is plotted on

Unique line \( l_u \): \( \gamma = 1 \)  

(4)

The second criterion is defined for a Dmap in Fig. 2 (b): if a Dmap has no EPP (no transient phenomenon) and periodic points are dense then \((\alpha, \gamma)\) is plotted on

Dense line \( l_d \): \( \alpha = 1 \)  

(5)

The third criterion is defined for a Dmap in Fig. 2 (d): if all the PEOs are fixed points then \((\alpha, \gamma)\) is plotted on

Fixed point curve \( C_f \): \( \gamma = \frac{1}{N\alpha} \)  

(6)

In the feature plane, we note three end points. If the Dmap has only one fixed point such as Fig. 2(a) then it is plotted at the left top point \((\alpha, \gamma) = (1/N, 1)\). If the Dmap has \( N \) fixed points and has no EPPs such as Fig. 2(b) then it is plotted at the right bottom point \((\alpha, \gamma) = (1, 1/N)\). If the Dmap has one PEO with period \( N \) and has no EPP such as an M-sequence in Fig. 2(c) then it is plotted at the right top corner \((\alpha, \gamma) = (1, 1)\).

3. Cellular automata and digital return maps

We consider ECAs on the ring of \( M \) cells. Let \( x_i^t \in \{0, 1\} \equiv B \) be the binary state of the \( i \)-th cell at discrete time \( t \) where \( i = 1 \sim M \). The time evolution of \( x_i^t \) is governed by a Boolean function of \( x_i^t \) and its closest neighbors:

\[ x_{i+1}^{t+1} = F_i(x_{i-1}^t, x_i^t, x_{i+1}^t), \quad i = 1 \sim M \]  

(7)

![Figure 2: Typical Dmaps for \( N = 16 \). (a) One fixed point. \((\alpha, \gamma) = (1/16, 1)\). (b) 16 fixed points and no EPPs. \((\alpha, \gamma) = (1, 1/16)\). (c) One PEO with period 16 (M-sequence) and no EPP. \((\alpha, \gamma) = (1, 1)\). (d) Three fixed points. \((\alpha, \gamma) = (3/16, 1/3)\). (e) \( \alpha-\gamma \) feature plane. \( l_u \): Unique line. \( l_d \): Dense line. \( C_f \): Fixed point curve.](image-url)
Figure 3: Examples of spatiotemporal patterns. (a) RN30 (b) RN90 (c) RN36 (d) RN102

The spatiotemporal pattern in Fig. 3(d) corresponds to a Dmap having 256 fixed points

$$x^{t+1} = F_D(x^t), \quad x^t \equiv (x^t_1, \ldots, x^t_M) \in B^M$$

Since $B^M$ is equivalent to a set of $2^M$ points $I_D \equiv \{C_1, \ldots, C_{2^M}\}$, $F_D$ is equivalent to the Dmap from $I_D$ to itself.

We focus on the case $M = 8$ ($N = 2^M = 256$) in this paper. Fig. 4(a) shows a Dmap and its PEO with period 8 corresponding to the spatiotemporal pattern in Fig. 3(a). The Dmap has 5 PEOs as shown in the histogram. This Dmap is characterized by $(\alpha, \gamma) = (0.2, 0.64)$ and plotted at cross A in the feature plane in Fig. 5. The cross A corresponds to the following 4 rules.

$$\text{RS-A} = \{30, 86, 135, 149\}$$

Fig. 4(b) shows a Dmap corresponding to the spatiotemporal pattern in Fig. 3(b). The steady state is one fixed point only. This Dmap is characterized by $(\alpha, \gamma) = (0.004, 1)$ and plotted at cross B in the feature plane in Fig. 5. The cross B corresponds to the following 12 rules.

$$\text{RS-B} = \{0, 8, 60, 64, 90, 102, 153, 165, 195, 239, 253, 255\}$$

Fig. 4(c) shows a Dmap corresponding to the spatiotemporal pattern in Fig. 3(c). This Dmap has 21 fixed points. This Dmap is characterized by $(\alpha, \gamma) = (0.08, 0.05)$ and plotted at cross C in the feature plane in Fig. 5. The cross C corresponds to the following 8 rules.

$$\text{RS-C} = \{36, 44, 72, 100, 203, 217, 219, 237\}$$

4. Conclusions

In order to analyze steady states of Dmaps, two simple feature quantities and feature plane are presented in this paper. Using the feature quantities, a simple class of ECAs are analyzed. Future problems include more detailed analysis of ECAs, analysis of mixed rule cellular automat[10], and engineering applications such as information compressions[6].

References

Figure 4: Examples of Dmaps from ECAs. (a) RN30, \((\alpha, \gamma) = (0.2, 0.64)\). Red points denote PEP. (b) RN90, \((\alpha, \gamma) = (0.004, 1)\). (c) RN36, \((\alpha, \gamma) = (0.08, 0.05)\).

Figure 5: \(\alpha - \gamma\) feature plane. \(l_u\): Unique line. \(l_d\): Dense line. \(c_f\): Fixed point curve.
A Variety of Super-Stable Periodic Orbits in a Simple Dynamical System with Integrate-and-Fire Switching

Risa Takahashi† and Toshimichi Saito‡

†Department of Electronics and Electrical Engineering, Hosei University
3–7–2 Kajino-cho, Koganei, Tokyo 184-8584, Japan
Emails: risa.takahashi.3j@stu.hosei.ac.jp; tsaito@hosei.ac.jp

Abstract—This paper studies dynamics of a simple switched dynamical system. Repeating integrate-and-fire behavior between a constant threshold and piecewise linear periodic base signal, the system can exhibit a variety of chaotic and super-stable periodic orbits. The dynamics is simplified into a piecewise linear return map. Using the return map, parameter conditions of various super-stable periodic orbits are analyzed precisely.

1. Introduction

This paper studies a variety of super-stable periodic orbits in a simple switched dynamical system (SDS) [1]-[3]. The dynamics of SDS is based on integrate-and-fire behavior between a constant threshold signal and periodic base signal. Especially, we consider the case where the base signal given by addition of the first triangular signal with period \( T \) and the second triangular signal with period \( T/M \). If the base signal is given by either the first of second signal only, the SDS exhibits chaotic orbit characterized by positive Lyapunov exponent. However, if the base signal is given by both the first and second signal, the SDS exhibits a variety of super-stable periodic orbit (SSPO) such that almost all initial state fall rapidly into the SSPO.

The dynamics of SDS is integrated into a piecewise linear return map. Based on the map, we present simple and systematic calculation methods of parameter regions of the SSPOs. We then clarify that, a variety of SSPOs exists or co-exist in the parameter space. Although discussion of this paper is based on theory-based numerical experiments, we have prepared laboratory measurements of typical phenomena for the final version.

The SDS is inspired by a simple integrate-and-fire type spiking neuron model [1]-[6]. Analysis results of simple neuron models have contributed to consideration of neural information processing function [7] [8]. Engineering applications of such systems are many, including signal/image processing and ultra wide band communication [9] [10]. Analysis of such systems is important not only as a nonlinear dynamical system but also for engineering applications.

It should be noted that our previous paper [3] has discussed SSPOs, however has not discussed parameter regions of various SSPOs and their calculation method.

2. Circuit Model

Figures 1 and 2 show a circuit model and dynamics of the SDS, respectively. The capacitor voltage \( v \) rises to the threshold \( V_T \) with slope \( s \). As \( v \) reaches the threshold then the SDS outputs a spike \( Y(t) = E \) and \( v \) reset to the periodic base signal \( B(t) \) with period \( T \). Repeating this integrate-and-fire behavior, the SDS outputs spike-train \( Y(t) \). For simplicity, the inner resistors are ignored \((r_1 \to \infty, r_2 \to 0)\) and the switching is assumed to be ideal: \( v_1 \) is reset instantaneously without delay. The dynamics is described by

\[
\begin{align*}
\frac{dv}{dt} &= I, \quad Y(t) = -E \quad \text{for } v(t) < V_T \\
v(t+\tau) &= B(t), \quad Y(t) = E \quad \text{if } v(t) = V_T \\
B(t) &= K_1 B_1(t) + K_2 B_3(t) + E_0, \quad B_1(t + T) = B_1(t)
\end{align*}
\]

where \( B(t) < V_T \). Using dimensionless variables and parameters:

\[
\begin{align*}
\tau &= \frac{T}{T} \quad x = \frac{v}{V_T}, \quad \dot{x} = \frac{dx}{d\tau}, \quad y = \frac{Y + E}{2} \\
k_1 &= K_1/\sqrt{V_T}, \quad k_3 = K_3/\sqrt{V_T}, \quad d = E_0/\sqrt{V_T}, \quad s = \frac{IT}{CV_T}
\end{align*}
\]

Eqs. (1)-(4) are transformed into

\[
\begin{align*}
x' &= y \\
x(\tau_*) &= b(\tau_*) \quad \text{for } x < 1 \\
x(\tau_*) &= b(\tau) \quad \text{if } x(\tau) = 1
\end{align*}
\]

where \( \tau_* \equiv dx/d\tau \). The base signal is characterized by parameters \( A \) and \( d \). For simplicity, we assume

\[
2 < A < 4, \quad 0 < d < 0.5
\]

In this paper, we consider three cases of the \( b(\tau) \).

Case 1: The first component only \((k_1 \neq 0, k_3 = 0)\)

Case 2: The second component only \((k_1 = 0, k_3 \neq 0)\)

Case 3: Two inputs \((k_1 \neq 0, k_3 \neq 0)\)

It goes without saying that the theorem of superposition is not valid in this nonlinear system.

- 226 -
3. Experiments

In order to confirm typical phenomena, we have fabricated a breadboard prototype of the BN. Figure 3 shows typical phenomena. The BN exhibits chaos for $B(t) = B_1(t)$ (first component only) or $B(t) = B_3(t)$ (second component only). However, if $B(t) = B_1(t) + B_3(t)$ then the BN exhibits periodic waveform as shown in Fig. 3(c). That is, chaotic behavior of each BN can be changed into periodic behavior by the two inputs.

4. Spike-phase Map

In order to analyze the dynamics, we derive a SPmap of the SDS. Let $\tau_n$ denote the $n$-th spike position. Since $\tau_{n+1}$ is determined by $\tau_n$, we can define the Smap.

$$\tau_{n+1} = \tau_n + (1 - b(\tau_n))/s \equiv F(\tau_n)$$ (8)

Since $F_1(\tau + 1) = F_1(\tau) + 1$ is satisfied, we introduce the phase variable $\theta_1(n) = \tau_1 \mod 1$. Using this, we can define the Pmap as shown in Fig. 4:

$$\theta_{n+1} = f(\theta_n) \equiv F(\theta_n) \mod 1$$ (9)

Substituting $k_1 = 0$ and $a_0 = 0$ into Eq. (6), we obtain the Pmap for the first component. Substituting $k_1 = 0$ into Eq. (6), we obtain the Pmap for the second component.

$$f(\theta_n) = \begin{cases} a\theta & \text{for } -d < \theta < d \\ -a\theta + (1 + a)/2 & \text{for } d < \theta < 1 - d \end{cases}$$ (10)

The shape of the Pmap depends on the shape of $b(\tau)$. As the parameter varies, the shape of Pmap varies and SDS can exhibit various spike-trains. Since the base signal is piecewise linear, the maps are also piecewise linear and precise numerical analysis is possible.

In Case 1 and 2, the Pmap exhibits chaos as shown in Fig. 5(a) and (b). In Case 3, the Pmap exhibits super-stable periodic orbit (SSPO) as shown in Fig. 6. That is, chaotic behavior of each SDS whose base signal is given by single component can be changed into periodic behavior by the two inputs.
5. Analysis

First, we give several basic definitions. A point \( \theta_f \in I \) is said to be a 1-periodic point or fixed point if \( F(\theta_f) = \tau_f \). A point \( \theta_p \in I \) is said to be a k-periodic point if \( F^k(\theta_p) = \theta_p \) and \( F^l(\theta_p) \neq \theta_p \) for \( 1 \leq l \leq k \) where \( F^k \) is the k-fold composition of \( F \) and \( k \geq 2 \).

Now we consider the periodic behavior in Case 3. Fig. 7 shows several Pmaps. Corresponding SDS exhibits periodic waveform with period 2. Note that the Pmap has four segments with zero-slope (\( s_1 \) to \( s_4 \)) and exhibit super-stable periodic orbits (SSPO). Let points \( p_1 \) to \( p_4 \) denote images of \( s_1 \) to \( s_4 \) as shown in Fig. 6.

We explain an outline of the analysis. First, since the Pmap includes four segments with slope-zero and other segments are expanding, a trajectory started from one of the four points \( p_1 \) to \( p_4 \) must return to either of the four points. That is, it is sufficient to use one of the four points as an initial values. If a trajectory started from \( p_i \) (\( i = 1 \sim 4 \)) returns to \( p_i \) then the orbit is SSPO. If a trajectory started from \( p_i \) falls into \( p_j (j \neq i) \) then orbit is transient to some SSPO.

Fig. 7 shows typical Pmaps of SSPOs with period 2. Fig. 8 shows parameter regions for SSPO with period 2: a basic
results of the bifurcation analysis in the $a_0 - a$ plane. It is
derived by the following 3 steps.

Step 1: Select the dc component $a_0$ and the slope $a$ as
control parameters.

Step 2: Use $p_i$ ($i = 1 \sim 4$) as an initial values.

Step 3: If a trajectory return to $p_i$ by twice then plot to the
$a_0 - a$ plane.

After the step is terminated, we obtain parameter regions
for SSPO with period 2.

6. Conclusions

We have studied a variety of super-stable periodic or-
bits in a simple dynamical system with integrate-and fire
switching. Using Pmaps, we have calculated parameter re-
gions of various SSPOs. The future problem includes de-
tailed analysis of typical bifurcation phenomena and appli-
cation to spike-based engineering systems.

References

[1] H. Torikai and T. Saito, Return map quantization from
an integrate-and-fire model with two periodic inputs,
IEICE Trans. Fundamentals, E82-A, 7, pp. 1336-1343,
1999.

pulse-coupled spiking neurons with a triangular base
signal, Neural Networks, 18, pp. 523-531, 2005.

various super-stable periodic spike-trains in bifur-
cating neuron with two triangular inputs, Proc. of
NOLTA, pp. 177-180, 2015.

[4] R. Perez and L. Glass, Bistability, period doubling bi-
furcations and chaos in a periodically forced oscilla-


realization of arbitrary one-dimensional maps, IEEE

[7] E. M. Izhikevich, Simple model of spiking neurons,
IEEE Trans. Neural Networks, 14, 6, pp. 1569-1572,
2003.

[8] N. Ning, H. Kejie and S. Luping, Artificial neuron
with somatic and axonal computation units: mathe-
matical and neuromorphic models of persistent firing
neurons, Proc. of IEEE-INNS/IJCNN, pp. 473-479,
2012.

[9] S. R. Campbell, D. Wang, and C. Jayaprakash, Syn-
chrony and desynchrony in integrate-and fire oscilla-

and A. R. Volkovskii, Digital communication using
chaotic-pulse-position modulation. IEEE Trans. Cir-
Homomorphisms from the Logistic Map to the Quadratic Maps over \( \mathbb{Z}_p \)

Takeru Miyazaki,\(^{1,2}\) Shunsuke Araki,\(^1\) Satoshi Uehara\(^3\) and Yasuyuki Nogami\(^4\)

\(^1\)Faculty of Environmental Engineering, the University of Kitakyushu, 1–1 Hibikino, Wakamatsu-ku, Kitakyushu City, Fukuoka 808-0135, Japan
\(^2\)Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680–4 Kawazu, Iizuka City, Fukuoka 820-8502, Japan
\(^3\) Faculty of Engineering, Okayama University, 3–1–1 Tsushimanaka, Kita-ku, Okayama City, Okayama 700-8530, Japan
\(^4\) Email: miyazaki@kitakyu-u.a.jp

Abstract—We have been designing a pseudorandom number generator with iteration maps, and studying characteristic properties of the logistic maps over prime fields, where a prime field means modular arithmetics with a prime number \( p \). The maps behave like quadratic maps over prime fields. In this present paper, we prove a homomorphic relation between the original logistic map and the quadratic map. We also prove that the number of structures constructed by these maps is only \((p + 1)/2\) and all of these maps correspond to the logistic maps over prime field by automorphism.

1. Introduction

Logistic map is one of the most famous chaotic maps[1]. It can produce a long and unpredicted sequence by an iterative mapping. In the implementation accurately for the computers, since the number of precision for the mapped value is twice or three times as many bits as one of the input value, we calculate the iterative maps on a finite precision arithmetic. Many studies such as [2] have been using floating points to implement the iterative map. On the contrary, we have been studying the logistic map over integers[3, 4]. In these methods, all calculations become integer values with fixed precisions and rounded fragments. Though they are good for an implementing on computers, they are also hard to analyze them theoretically. We also presented the logistic map over prime field[5, 6, 7], which is based on the modulus calculation with a prime number \( p \), so that the elements of input and mapped values by this map become integer values in \([0, p - 1]\). This method is suitable for theoretical analyses. We estimate that one of the reasons of it comes from no truncation part in the calculation. For example, Tsuchiya and Nogami have analyzed a period of the loop in the generated sequence by this map under specific conditions[8]. We have proved automorphic relations on the maps with two distinct control parameters[9]. We also found that the logistic map over prime field is included in the one of the quadratic maps over the prime fields. Some studies have already considered the properties of the map over modulus an integer. For example, Knuth has proved the longest period of the sequence generated by the map[10]. However, his proof cannot apply to the maps over the prime field. Hence, we are interested in the properties of each sequence generated by the map and their variations.

In this present paper, we propose a homomorphic relation between the original logistic map and the quadratic maps over the prime fields. Since the quadratic maps include all of the logistic maps over prime fields, it also explains the reason why the automorphic relations are occurred. We also prove that the number of structures constructed by these quadratic maps is only \((p + 1)/2\) and all of these maps correspond to the logistic maps over prime field by automorphism.

2. Preparation

2.1. Logistic Maps

The logistic map over the real domain is given by

\[
\text{LM}_R(r) = \mu r(1 - r),
\]

where \( r \) is a real number in \([0, 1]\), and \( \mu \) is a control parameter. Let \( r_i \) be an input, where \( i \) acts as the discrete time. The iterative mapping for Eq. 1 can be written as follows:

\[
r_{i+1} = \mu r_i(1 - r_i).
\]

We derive the logistic map over integers from Eq. 1. Let \( n \) be the precision for elements. By defining \( \bar{r} = 2^n r \) and \( \text{LM}_R^{(n)}(\bar{r}) = 2^n \text{LM}_R(r) \), Eq. 1 can be transformed into

\[
\text{LM}_R^{(n)}(\bar{r}) = \mu \bar{r}(2^n - \bar{r})/2^n.
\]

Let \( X \) be the integer part of \( \bar{r} \) and \( [\bar{r}] \) be the floor function, which outputs the integer part of \( \bar{r} \). We define a function for the logistic map over integers as

\[
\text{LM}_R^{(n)}(X) = [\mu X(2^n - X)/2^n].
\]

where \( X \in [0, 2^n] \). Using a method similar to that demonstrated for Eq. 2, the iterative mapping for Eq. 3 is

\[
x_{i+1} = [\mu x_i(2^n - x_i)/2^n].
\]
2.2. Logistic Map over Prime Fields

We define a logistic map over prime field $\text{LM}_{Z_p[X]}(X)$ as follows.

**Definition 1.** Let $p$ be an odd prime, $Z_p$ a prime field modulo $p$, and $X$ an element in $Z_p$. Then we define the logistic map over prime field $\text{LM}_{Z_p[X]}(X)$ as

$$\text{LM}_{Z_p[X]}(X) = \frac{\mu_p X(p - 1 - X)}{p - 1} \text{ mod } p, \quad (4)$$

where $\mu_p$ is a control parameter with $\mu_p \in [1, p - 1]$.

According to the number of fixed point, we can classify the map more efficiently than Eq. 4.

**Lemma 1.**

$$\text{LM}_{Z_p[X]}(X) = \mu_p X(X + 1) \text{ mod } p.$$  

**(Proof)**

By using Eq. 4, we obtained the following equation:

$$\text{LM}_{Z_p[X]}(X) = \frac{\mu_p X(p - 1 - X)}{p - 1} \equiv \frac{\mu_p X(-1 - X)}{p} \equiv \mu_p X(X + 1) \text{ (mod p).} \quad \square$$

2.3. Quadratic Map over Prime Fields

A map $f(x) = ax^2 + bx + c$ with $a_2 \neq 0$ is called as a quadratic map. We now define a quadratic map over prime field $\text{QM}_{Z_p}(X)$ as follows.

**Definition 2.** Let $p$ be an odd prime, $Z_p$ a prime field modulo $p$, and $A, B, C$ and $X$ four elements in $Z_p$, where $A \neq 0$ (mod $p$). Then we define the quadratic map over prime field $\text{QM}_{Z_p}(X)$ as

$$\text{QM}_{Z_p}(X) = AX^2 + BX + C \text{ mod } p. \quad (5)$$

Then, we prove the next lemma which means that $\text{LM}_{Z_p[X]}(X)$ is included in a part of $\text{QM}_{Z_p}(X)$.

**Lemma 2.** $\text{LM}_{Z_p[X]}(X)$ is equal to $\text{QM}_{Z_p}(X)$ when $A = B = \mu_p$ and $C = 0$.

**(Proof)**

If $A, B$ and $C$ satisfy these conditions, then

$$\text{QM}_{Z_p}(X) \equiv \mu_p X(X + 1) \equiv \text{LM}_{Z_p[X]}(X) \text{ (mod p).} \quad \square$$

2.4. Sequences generated by $\text{LM}_{Z_p[X]}(X)$ and $\text{QM}_{Z_p}(X)$

Since the ranges of both input value and mapped one are the same as $[0, p - 1]$, $\text{LM}_{Z_p[X]}(X)$ can calculate iteratively for any times. Let $X_0$ be an element in $Z_p$ satisfying

$$X_{i+1} = \text{LM}_{Z_p[X]}(X_i), \quad i = 0, 1, \cdots$$

(a) Case 1: two fixed points

$$X_{i+1} = X_i^2 + X_i \mod 11$$

(b) Case 2: one fixed point

$$X_{i+1} = X_i^2 + X_i \mod 11$$

(c) Case 3: no fixed point

$$X_{i+1} = X_i^2 + X_i \mod 11$$

Figure 1: Three trajectories of $\text{QM}_{Z_p}(X)$ with $p = 11$, where $A = 1, B = 1$ and $C$ of (a),(b) and (c) is 2,0 and 1, respectively

and let $S_L$ a generated sequence by $\text{LM}_{Z_p[X]}(X)$ with an initial value $X_0$ as

$$S_L = (X_0, X_1, \cdots)$$

By using the same manner, a generated sequence $S_Q$ by $\text{QM}_{Z_p}(X)$ also is described by

$$S_Q = (X_0, X_1, \cdots), \quad X_{i+1} = \text{QM}_{Z_p}(X_i), \quad i = 0, 1, \cdots$$

2.5. Number of Fixed Points on $\text{QM}_{Z_p}(X)$

Let $X$ be an element in $Z_p$. Then, $X$ is a fixed point on $\text{QM}_{Z_p}(X)$ if and only if $X$ satisfies the following equation:

$$X = \text{QM}_{Z_p}(X). \quad (6)$$

According to the number of fixed point, we can classify the maps. For example, Fig. 1 illustrates three trajectories with three cases (a), (b) and (c). Though all of them are based on $Z_p$ with the same $p = 11$, the number of fixed points in them are different each other.

By using Eq. 6, we obtained the following equation:

$$AX^2 + (B - 1)X + C \equiv 0 \text{ (mod } p).$$

Let $D$ be a discriminant of $\text{QM}_{Z_p}(X)$, where $D = (B - 1)^2 - 4AC \text{ mod } p$. Then, the number of fixed points on $\text{QM}_{Z_p}(X)$ is defined by $D$ as the following three cases.

- **Case 1**: $D$ is a quadratic residue.
  
  There are two distinct and non-zero values $E$ and $-E$ satisfying $(\pm E)^2 \equiv D \text{ (mod } p)$. Then, Eq. 6 has just two distinct solutions as $X \equiv (1 - B \pm E)/2A \text{ (mod } p)$. Therefore, the number of the fixed points are also two.

- **Case 2**: $D$ is zero.
  
  Equation 6 has only one solution as $X \equiv (1 - B)/2A \text{ (mod } p)$. Therefore, the number of the fixed points is also just one.
• Case 3: $D$ is a quadratic non-residue.

Since Eq. 6 has no solution, the number of the fixed points is zero.

Since the number of elements with quadratic residues is equal to one with quadratic non-residues, the number of $D$ satisfying Cases 1 and 3 are also the same as $(p - 1)/2$, and there is just only one $D = 0$ satisfying Case 2.

3. Homomorphic Relation between $LM_{Z_2[q]}(X)$ and $QM_{Z_2}(X)$

In this section, we prove the next two theorems which propose homomorphic relations from $LM_{R}(r)$ with rational numbers to $QM_{Z_2}(X)$ and $LM_{Z_2[q]}(X)$.

**Theorem 1.** Let $q, \mu_q$ be rational numbers with $q$ in the closed interval $[0, 1]$ and $\mu_q$ in the half-opened interval $(0, 4]$. If parameters satisfy Case 1 in Section 2.5, then there exists a homomorphic relation from $LM_{R}(q) = \mu_q q(1 - q)$ to two of $QM_{Z_2}(X)$ defined by homomorphic maps $Hom(q)$ such that

$$Hom(q) = Sq + T \pmod{p},$$

where

$$S = \frac{-1 \pm E}{A}, T = \frac{1 - B \mp E}{2A}, \text{ and } \mu_q = -AS \pmod{p}. \ (7)$$

**Proof**

We prove that $Hom(LM_{R}(q)) \equiv QM_{Z_2}(Hom(q)) \pmod{p}$.

$$Hom(LM_{R}(q)) \equiv S\mu_q q(1 - q) + T \pmod{p},$$

$$\equiv AS^2 q^2 - AS^2 q + T \pmod{p},$$

and

$$QM_{Z_2}(Hom(q)) \equiv A(Sq + T)^2 + B(Sq + T) + C \pmod{p},$$

$$\equiv AS^2 q^2 - AT^2 + (B - 1)T + C \pmod{p}.$$ 

Since $AT^2 + (B - 1)T + C \equiv 0 \pmod{p}$, we can get

$$QM_{Z_2}(Hom(q)) \equiv AS^2 q^2 - AS^2 q + T \pmod{p}.$$ 

Therefore,

$$Hom(LM_{R}(q)) \equiv QM_{Z_2}(Hom(q)) \pmod{p}. \ \Box$$

We can also prove the next Corollary by the same way of Theorem 1 with $E = 0$.

**Corollary 1.** In Theorem 1, if parameters satisfy Case 2 instead of Case 1, then there also exists a homomorphic relation from $LM_{R}(q) = \mu_q q(1 - q)$ to just one of $QM_{Z_2}(X)$ by homomorphic map $Hom(q)$ such that $Hom(q) = Sq + T \pmod{p}$, where $S \equiv (-1)/A \pmod{p}$, $T \equiv (1 - B)/2A \pmod{p}$, and $\mu_q \equiv -AS \pmod{p}. \ \Box$

Next, we consider the following situation that $LM_{R}(q)$ with one control parameter $\mu_q$ has homomorphic relations with two distinct maps $Q_1(q)$ and $Q_2(q)$ which are included in $QM_{Z_2}(X)$ by using two homomorphic maps $Hom_1(q)$ and $Hom_2(q)$, respectively. Let $X_1$ and $X_2$ be elements in $Z_p$ such that

$$X_1 = Hom_1(q), \text{ and } X_2 = Hom_2(q).$$

Then, we prove the next lemma.

**Lemma 3.** There is an automorphic relation between $Q_1(q)$ and $Q_2(q)$.

**Proof**

Let $q_p \equiv q \pmod{p}$. Then, $Hom_1(q_p)$ and $Hom_2(q_p)$ are two distinct one-to-one mappings on $Z_p$. Let $Hom_1^{-1}(X)$ be an inverse map of $Hom_1(q)$, such that $q_p = Hom_1^{-1}(X_1)$. Then, $Hom_2^{-1}(X)$ is also a one-to-one mapping. Therefore, $X_1$ and $X_2$ can convert each other by using a one-to-one mapping

$$X_2 = Hom_2(Hom_1^{-1}(X_1)).$$

This means that $X_1$ and $X_2$ have an isomorphic relation each other. Since $X_1$ and $X_2$ are in $Z_p$, there is an automorphic relation between $X_1$ and $X_2$. \ \Box

Figure 2 demonstrates an example of the homomorphic relations from $LM_{R}(q) = 5q(1 - q)$ to $QM_{Z_2}(X) = 2X^2 + 2X + 5 \pmod{11}$ and $QM_{Z_2}(X) = 3X^2 + 4X + 4 \pmod{11}$, with homomorphic functions $Hom(q) = 3q + 9 \pmod{11}$ and $Hom(q) = 10q + 2 \pmod{11}$, respectively. Then, these two quadratic maps have an automorphic relation, so that these two trajectories are the same structures.

**Theorem 2.** $LM_{R}(q)$ with $\mu_q = A$ and $(2 - A)$ satisfy a homomorphic relation with $LM_{R,(p,\mu_p)}(X)$ with $\mu_p = A$ and $\mu_p = (p + 2 - A)$ where $A = 3, 4, \cdots, p - 1$.

**Proof**
By Lemma 2, if $\text{QM}_{Z_p}(X)$ is equal to $\text{LM}_{Z_p[A]}(X)$, $A = B$ and $C = 0$. Hence, we can get

$$D^2 = (B - 1)^2 = (A - 1)^2, \ E = \pm (A - 1).$$

When $E = A - 1$, $S$, $T$, and $\mu_p$ are defined, Eqs. 7 become

$$S = -1, \ T = 0, \ \mu_q = A.$$

This means that there is a homomorphic relation from $\text{LM}_{R}(q) = Aq(1 - q)$ to $\text{LM}_{Z_p[A]}(X) = AX(X + 1)$ mod $p$ by using $\text{Hom}(q) = -q$ mod $p$.

When $E = -(A - 1)$, $S$, $T$, and $\mu_p$ become

$$S = \frac{A - 2}{A}, \ T = \frac{1 - A}{A}, \ \mu_q = 2 - A.$$

This means that there is a homomorphic relation from $\text{LM}_{R}(q) = Aq(1 - q)$ to $\text{LM}_{Z_p[2-A]}(X) = (2 - A)X(X + 1)$ mod $p$ by using $\text{Hom}(q) = \frac{(4 - 2q + 1 - A)}{A}$ mod $p$.

If $A \geq 3$, $A$ and $(2 - A)$ are distinct and non-zero element. Therefore, two maps $\text{LM}_{Z_p[A]}(X)$ and $\text{LM}_{Z_p[2-A]}(X)$ are corresponding to the same map $\text{LM}_{R}(q) = Aq(1 - q)$.

Moreover, we can also see that two maps $\text{LM}_{Z_p[A]}(X)$ and $\text{LM}_{Z_p[2-A]}(X)$ are corresponding to the same map $\text{LM}_{R}(q) = A'q(1 - q)$, where $A' = 2 - A$ (mod $p$). Since $2 - A' = A$, there are two homomorphic relations from one map $\text{LM}_{R}(q) = (2 - A)q(1 - q)$ to $\text{LM}_{Z_p[2-A]}(X)$ and $\text{LM}_{Z_p[A]}(X)$.

Therefore, two maps $\text{LM}_{R}(q) = Aq(1 - q)$ and $\text{LM}_{R}(q) = (2 - A)q(1 - q)$ have homomorphic relations to two maps $\text{LM}_{Z_p[A]}(X)$ and $\text{LM}_{Z_p[2-A]}(X)$.

By using Theorem 2, the number of structures on the sequences generated by $\text{LM}_{Z_p[A]}(X)$ with $\mu_p \in [3, p - 1]$ becomes $(p - 3)/2$. Since the structures by $\text{LM}_{Z_p[A]}(X)$ with $\mu_p = 1$ and $\mu_p = 2$ are not the same as them, there are only $(p + 1)/2$ of structures by $\text{LM}_{Z_p[A]}(X)$. By using Theorem 1 and Lemma 3, we can see that all of $\text{QM}_{Z_p}(X)$ with Cases 1 and 2 in Section 2.5 have an automorphic relation with one of $\text{LM}_{Z_p[A]}(X)$. Therefore, the number of structures on the sequences generated by these quadratic maps is just $(p + 1)/2$, too.

4. Conclusion

In this present paper, we have discussed a few homomorphic relations from $\text{LM}_{R}(q)$ into $\text{QM}_{Z_p}(X)$. Since $\text{LM}_{Z_p[A]}(X)$ has an automorphic relation with each $\text{QM}_{Z_p}(X)$ with Cases 1 and 2, and the number of structures for all trajectories of $\text{LM}_{Z_p[A]}(X)$ is $(p + 1)/2$, the number of structures for trajectories generated by the quadratic maps is also just $(p + 1)/2$.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number JP16H01723.
A Biased Distribution of Truncated Values in the Logistic Map over Integers

Hideyuki Muraoka\textsuperscript{1}, Shunsuke Araki\textsuperscript{2}, Takeru Miyazaki\textsuperscript{3}, Satoshi Uehara\textsuperscript{4} and Ken’ichi Kakizaki\textsuperscript{5}

\textsuperscript{1}Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
Email: q794053h@mail.kyutech.jp
\textsuperscript{2}Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology
680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan
\textsuperscript{3}Faculty of Environmental Engineering, the University of Kitakyushu
1-1 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0135, Japan

Abstract—In the presented paper, we will discuss a distribution of truncated values in the calculation process of the logistic map over integers, where these values are outputs by the logistic map modulo a power of two. Although originally the logistic map has only two solutions for an output, we will observe that there are many solutions for this modulo-based logistic map by some numerical experiments and then will derive a condition that many inputs are mapped into zero on the logistic map. Additionally, we will show that the input range is separated into eight blocks which have the same pattern of outputs. Conclusively, we will mention that the resolution of the logistic map modulo the power of two reduces four bits at least.

1. Introduction

One purpose of our research is to design a pseudorandom number generator which is one of the most important fundamental technologies for information security. For a simple construction, we focused on the logistic map, popularized in [1], for generating pseudorandom numbers. The logistic map must be computed on finite precision, because a computer has only restricted memories. In [2], Phatak and Rao proposed a pseudorandom number generator using the logistic map whose values are represented in the floating-point format. In [3], Pareek, Patidar and Sud used the logistic map represented in the floating-point format for their image encryption method. Expecting a simple format and a fast computation, we have been studying on the logistic map over integers. Our group showed some results about characteristic behaviors for the logistic map over integers [4].

We configured a basic model of the pseudorandom number generator with an iterative map consisting of three basic operations, namely a bit extraction, iterations without extracting bits and a renewal of a parameter. Recently, Persohn and Povinelli pointed out some cyclic behaviors of the logistic map by floating-point arithmetic in [5]. However, before their work, in [6, 7] we showed that the pseudorandom number generator employing a basic model with the logistic map with finite precision can generate pseudorandom number outputs passing the NIST statistical test suites[8]. In [9, 10], we got an important result about bit positions in order to extract bits from outputs generated by the logistic map over integers. Although the result is only for the control parameter 4, we showed 0/1 occurrence rate per bit of values by the map theoretically. Note that the values are in the former step for the process which is to cut off low-order bits in order to fit in required bit length as the output. This is one of the most important results, because it indicates a suitable bit position in order to extract bits as a part of pseudorandom bit sequences.

However, when these truncated bits regard as any integer, we have not investigated estimation for a distribution of these values yet. In this paper, we will define In this paper, we will define a modulo-based logistic map as the map generating these truncated values in calculation process of the logistic map over integers. For the modulo-based logistic map, we will discuss a distribution of outputs generated by the logistic map. Note that we will not treat the modulo-based logistic map as the iterative mapping in this

![Figure 1: Relation between Inputs and Outputs by LM\textsubscript{int}(X) and LM\textsubscript{R}(r)](image-url)
paper, because elements of a sequence by the logistic map over integers are given as inputs of the modulo-based logistic map in our assumed pseudorandom number generator.

2. The Logistic Map over Integers

The logistic map is given by

\[ \text{LM}(r) = \mu r (1 - r), \]

(1)

where \( r \) is a real number in the closed interval \([0, 1]\), and a control parameter \( \mu \) is a real number defined in the interval \([0, 4]\). In this paper, we will call this map the logistic map over real domain.

For implementation of this map on computers, these values must have finite precision. So, we would derive the logistic map over integers from Eq. (1). Let \( n \) be the precision for elements. By defining \( r = 2^n r \) and \( \text{LM}^{(n)}(r) = 2^n \text{LM}(r) \), Eq. (1) can be transformed into

\[ \text{LM}^{(n)}(r) = \mu 2^n (2^n r - r)/2^n. \]

Equation (2) becomes a basis for the logistic map over integers. Because an output of Eq. (2) is a real number in the closed interval \([0, 2^n]\) for each integer in the interval \([0, 2^n]\), the output must be mapped into an integer output in the interval \([0, 2^n]\) by some kind of rounding functions in the logistic map over integers. In this paper, we employ a floor function, which outputs the integer part of the logistic map over integers. In this paper, we will call this map the logistic map over integers with the control parameter \( \mu = 4 \), theoretically, where their estimated values are ones with double the arithmetic precision before truncating for the next inputs. This result gave suitable bit positions for a bit extraction which is one of basic operations of a pseudorandom number generator with iterative mapping, because elements of an output sequence by Eq. (4) are given as inputs of Eq. (6) in our assumed pseudorandom number generator.

Note that we defined this map only in order to represent low-order \( n \) bits of \( \mu X(2^n - X) \). We will not employ Eq. (6) as the iterative mapping, because elements of an output sequence by Eq. (4) are given as inputs of Eq. (6) in our assumed pseudorandom number generator.

Figure 1 shows relations between inputs and outputs for \( \text{LM}^{(n)}(X) \) and \( \text{LM}_{2^n}(X) \) with \( n = 7 \), and Figs. 2 and 3 illustrate relations between inputs and outputs for \( \text{LM}_{2^n}(X) \) with \( n = 7 \) and 8, respectively.

3. Distribution of Values by the Logistic Map Modulo \( 2^n \)

In [10], we proved 0-occurrence rate per bit of values by the logistic map over integers with the control parameter \( \mu = 4 \), theoretically, where their estimated values are ones with double the arithmetic precision before truncating for the next inputs. This result gave suitable bit positions for a bit extraction which is one of basic operations of a pseudorandom number generator with iterative mapping, because 0-occurrence rate is 50% for each of almost all bits except several low-order bits.

Values with high-order \( n \) bits show a distribution shown by Fig. 1. However, we do not have a distribution of values with low-order \( n \) bits which are calculated by \( 4X(2^n - X) \mod 2^n \). In order to obtain a trend of the distribution of \( 4X(2^n - X) \mod 2^n \), we got relations between inputs and outputs for some precisions. Figures 2 and 3 show relations between inputs and outputs by the logistic map modulo \( 2^n \) for representing these truncated values:

\[ \text{LM}_{2^n}(X) = 4X(2^n - X) \mod 2^n. \]

(6)
This theorem means that \(2^{n - 1}\) inputs satisfy \(\text{LM}_{Z_0}(X) = 0\). Additionally, we would give the following lemma for some inputs satisfying the above theorem.

**Lemma 2.** Let \(X = 2^\alpha\) be a value satisfying Theorem 1 for integers \(i\) in the closed interval \([0, n]\) and \(\alpha\) in the interval \([0, \lfloor \frac{n}{2} \rfloor]\). For \(i = n - 4\), it holds that \(\text{LM}_{Z_0}(X - \delta) = \text{LM}_{Z_0}(X + \delta)\), where \(\delta\) is any integer in the interval \([0, 2^n]\).

**Proof:** For \(X = 2^\alpha \pm \delta\), two results are shown as follows:

\[
\begin{align*}
\text{LM}_{Z_0}(X - \delta) &= 4(2^\alpha - \delta)(2^n - (2^\alpha - \delta)) \mod 2^n \\
&= -4(2^\alpha - \delta)^2 \mod 2^n \\
&= -2^{2+2}2^2 + 2^{3+2}\alpha\delta + 2^2\delta^2, \text{ and}
\end{align*}
\]

\[
\begin{align*}
\text{LM}_{Z_0}(X + \delta) &= 4(2^\alpha + \delta)(2^n - (2^\alpha + \delta)) \mod 2^n \\
&= -4(2^\alpha + \delta)^2 \mod 2^n \\
&= -2^{2+2}2^2 - 2^{3+2}\alpha\delta - 2^2\delta^2.
\end{align*}
\]

For \(\text{LM}_{Z_0}(X - \delta) = \text{LM}_{Z_0}(X + \delta)\), it holds that \(2^{i+3} \equiv 0, 2^{i-1} \mod 2^n\). If \(2^{i+3} \alpha \equiv 0 \mod 2^n\), \(i = n - 3\).

Note that the proper solution of \(2^{i+3} \alpha \equiv 0 \mod 2^n\) is \(i \geq n - 3\). Because \(2^\alpha\) for \(i = n - 3\) includes \(2^\alpha\) for \(j \geq n - 2\), we will use the solution in the rest of paper. For \(2^{i+3} \equiv 2^{i-1} \mod 2^n\), \(i = n - 4\). For even \(\alpha = 2\alpha'\), because \(2^{n-4}(2\alpha') = 2^{n-3}\alpha'\), the case of \(i = n - 4\) includes one of \(i = n - 3\).

According to the above lemma, we can obtain the following results about outputs by \(\text{LM}_{Z_0}(X)\).

For even \(\alpha = 2\alpha'\),

\[
\text{LM}_{Z_0}(2^{n-4}\alpha \pm \delta) = \text{LM}_{Z_0}(2^{n-3}\alpha' \pm \delta)
\]

\[
\begin{align*}
&= 4(2^{n-3}\alpha' \pm \delta) \cdot \\
&= 4(2^n - 2^{n-3}\alpha' \pm \delta) \mod 2^n \\
&= -4(2^n - 2^{n-3}\alpha' \pm \delta)^2 \mod 2^n \\
&= -4(2^{2n-8}\alpha'^2 \pm 2^{n-3}\alpha'\delta + 2^n\delta^2) \\
&= -4(2^{n-4}2\alpha'^2 \pm 2^n\alpha'\delta + 2^n\delta^2) \\
&= -4\delta^2.
\end{align*}
\]

where \(n \geq 4\).

For odd \(\alpha = 2\alpha' + 1\),

\[
\text{LM}_{Z_0}(2^{n-4}\alpha \pm \delta) = 4(2^{n-4}\alpha \pm \delta) \cdot \\
= 4(2^n - 2^{n-4}\alpha \pm \delta) \mod 2^n \\
= -4(2^n - 2^{n-4}\alpha \pm \delta)^2 \mod 2^n \\
= -4(2^{2n-9}\alpha^2 \pm 2^{n-3}\alpha\delta + 2^n\delta^2) \\
= -2^n\alpha\delta - 4\delta^2 \\
= -2^{n-1}(2\alpha' + 1)\delta - 4\delta^2 \\
= 2^{n-1}\delta - 4\delta^2,
\]

where \(n \geq 6\). These above results give the following theorem:
Theorem 3 Let \( \delta \) be an integer. For \( X = 2^{n-4}\alpha, 2^{n-4}\alpha \pm \delta \) gives the following outputs:

\[
\text{LM}_{Z_2}(X) = \begin{cases} 
-4\delta^2 \\ 2\alpha^2 - 4\delta^2 
\end{cases} \quad \text{for even } \alpha, \text{ and } \\
\text{for odd } \alpha.
\]

(7)

It is important that \( \text{LM}_{Z_2}(2^{n-2}\alpha \pm \delta) \) with odd and even \( \alpha \) differ from each other. Since \( \text{LM}_{Z_2}(2^{n-2}\alpha + \delta) = \text{LM}_{Z_2}(2^{n-2}\alpha + 2 + \delta) \) for even \( \alpha \) in the interval \([0, 2^n]\), as the largest block, there exist eight blocks which have the same pattern of outputs for each input range \([2^{n-4}\alpha, 2^{n-4}(\alpha + 2)]\).

Of course, since the interval \([2^{n-4}\alpha, 2^{n-4}(\alpha + 2)]\) for even \( \alpha \) includes \( 2^{n-4}(\alpha + 1) \), in each block there exists two patterns which are symmetrical centering on \( 2^{n-4}(\alpha + 1) \). Therefore, the input range \([0, 2^n]\) can be separated into 16 blocks which have a pattern. This consideration indicates that the resolution of \( \text{LM}_{Z_2}(X) \) reduces four bits at least.

4. Conclusion

In this paper, we discussed the logistic map modulo \( 2^n \):

\[ \text{LM}_{Z_2}(X) = 4X(2^n - X) \mod 2^n \]

which outputs a truncated bit sequence. We calculated the logistic map over integers for computing the next input. First, we observed two characteristic behaviors by a numerical experiment. Second, we proved that there exist many inputs satisfying \( \text{LM}_{Z_2}(X) = 0 \), concretely \( 2^\lceil\frac{n}{2}\rceil \), though in general the logistic maps have only two inputs. Last, we derived that there is an output pattern for inputs in the interval \([2^{n-3}\alpha, 2^{n-3}(\alpha + 1)]\), where \( \alpha \) in the interval \([0, 2^3 - 1]\), and the pattern appears eight times for inputs in \([0, 2^n]\). Additionally, there exists a pattern which the center is \( 2^{n-4}(2\alpha + 1) \). Therefore, we have an unfortunate result that the resolution of \( \text{LM}_{Z_2}(X) \) reduces four bits at least.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP16H01723.

References

The period of Chebyshev polynomial sequences modulo a prime power $p^k$

Daisaburo Yoshioka\textsuperscript{\dagger} and Kento Kawano\textsuperscript{\ddagger}

\textsuperscript{\dagger}Department of Computer and Information Sciences, Sojo University
4–22–1 Nishi-ku, Ikeda, Kumamoto 860-0082, Japan
\textsuperscript{\ddagger}Graduate School of Engineering, Sojo University
4–22–1 Nishi-ku, Ikeda, Kumamoto 860-0082, Japan

Abstract—A public-key cryptosystem based on Chebyshev polynomials has been recently proposed. In this paper, we give conditions on the degree of Chebyshev polynomials to be permutation polynomials modulo a prime power. We also derive the period of sequences generated by Chebyshev polynomials modulo a prime power.

1. Introduction

A polynomial over a finite ring is called a permutation polynomial if the mapping defined by the polynomial is one-to-one. Permutation polynomials have been used in cryptography, coding, and pseudorandom number generation. Rivest has shown a necessary and sufficient condition on coefficients of polynomials to be a permutation polynomial over a ring of integers modulo a power of two [1]. Umeno proved that Chebyshev polynomials of odd degree become permutation polynomials over the ring [2].

Taking advantage of the commutative property of Chebyshev polynomials in real field, a public key cryptosystem based on Chebyshev polynomials was firstly proposed [3], but soon broken [4]. In order to resist such attack, the definition of Chebyshev polynomials was expanded from real field to finite fields or finite rings [2], [5]. To analyze the security of the cryptosystem, several properties of sequences generated by iterating Chebyshev polynomials over a finite set (called Chebyshev polynomial sequences in the following discussions) have been investigated [6]–[10]. Indeed, it turns out that the cryptosystem employing Chebyshev polynomials over the integer ring of powers of two, even if efficient and practical, is unfortunately not secure [10]. The weakness of this algorithm is that Chebyshev polynomial sequences over the ring have regular periodicities. Therefore, it is important to clarify periodic properties of Chebyshev polynomials. In particular, it is necessary and sufficient condition on degree of Chebyshev polynomials to be a permutation polynomial over the ring of integers of powers of prime. We also derive a periodicity of Chebyshev polynomial sequences over the ring.

2. Chebyshev polynomial sequences modulo $p^k$

In this section, after we briefly give the definition and introduce some properties of Chebyshev polynomials, our new results will be presented.

2.1. Chebyshev polynomials modulo $p^k$

The Chebyshev polynomials of the first kind of degree $n$ are defined by the recurrence relation

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x), n = 2, 3, \cdots,$$

where $T_0(x) = 1$ and $T_1(x) = x$. The first few Chebyshev polynomials are $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x$, $T_4(x) = 8x^4 - 8x^2 + 1$.

One of the most important properties of Chebyshev polynomials is the semi-group property, that is, the composition of Chebyshev polynomials is also Chebyshev polynomials. In particular,

$$T_n(T_m(x)) = T_m(T_n(x)) = T_{nm}(x).$$

The commutative property allows us the construction of public-key cryptosystems. Up to linear transformations, the monomial $x^n$ that appears in the Diffie-Hellman key agreement protocol and the Chebyshev polynomials are the only classes of polynomials that satisfy the commutative property. Thus, the cryptosystems employing Chebyshev polynomials have been proposed by replacing $x^n$ with $T_n(x)$ [2],[3].

Let $\mathbb{Z}$ be the set of all integers and $p$ be a prime number. For a positive integer $k \geq 1$, we consider Chebyshev polynomials over the residue ring of integers $R = \mathbb{Z}/p^k\mathbb{Z}$. Namely,

$$y = T_n(x) \mod p^k.$$
We introduce the Dickson polynomial $D_n(x, a)$ of degree $n$ which is defined by

$$D_n(x, a) = \sum_{m=0}^{\lfloor n/2 \rfloor} \binom{n}{n-m} C_m(a)^m x^{n-2m},$$  \hspace{1cm} (4)

where $\lfloor \cdot \rfloor$ denotes greatest integer function. The first few Dickson polynomials are $D_0(x, a) = 2$, $D_1(x, a) = x$, $D_2(x, a) = x^2 - 2a$, $D_3(x, a) = x^3 - 3ax$, and so on. $D_n(x, a)$ over the ring $R = \mathbb{Z}/p^k\mathbb{Z}$ is a permutation polynomial if and only if the degree $n$ is relatively prime to both $p$ and $p^2 - 1$, where $a$ is a unit over the ring $R$ [11].

Dickson polynomials $D_n(x, a)$ are related to Chebyshev polynomials $T_n(x)$:

$$D_n(2x, 1) = 2T_n(x).$$  \hspace{1cm} (5)

**Lemma 1** Assume $a$ and $p$ are relatively prime. If $af(x)$ is a permutation polynomial over $R = \mathbb{Z}/p^k\mathbb{Z}$, then $f(x)$ is also a permutation polynomial over $R$.

**Proof:** Suppose that $f(x)$ is not a permutation polynomial over $R$, then there are some integers $x$ and $y$ such that $f(x) \equiv f(y) \mod p^k$, which implies $af(x) \equiv af(y) \mod p^k$. This contradicts the fact that $af(x)$ is a permutation polynomial over $R$. \hfill $\Box$

Since $2x$ is a permutation polynomial over $R$, $D_n(2x, 1)$ is also a permutation polynomial over $R$. Together with Lemma 1, we obtain the following result.

**Theorem 1** Let $p$ be an odd prime and $k > 1$ be an positive integer. A Chebyshev polynomial $T_n(x)$ is a permutation polynomial modulo $p^k$, if and only if $(n, p) = (n, p^2 - 1) = 1$, where $(a, b)$ denotes the greatest common divisor of two integers $a$ and $b$.

For example, when $p = 3$, $T_n(x)$ is a permutation polynomial modulo $3^k$ for any odd $n$ which is not multiples of three. Thus, we always assume that $n$ is an integer such that $(n, p) = (n, p^2 - 1) = 1$ hereafter.

The $i$-th iterate of $T_n(x)$ is denoted by

$$T_n^i(x) = T_n(T_n^{i-1}(x)) \mod p^k.$$

We can generate an integer periodic sequence by iterating (6) from an initial value $x$. The period $N$ is defined as the least positive integer such that

$$T_n^N(x) \equiv x \mod p^k.$$  \hspace{1cm} (7)

We now present examples. We can compute the sequence $\{T_n(x) \mod p^k\}_{i=0}^{N-1}$ until the period $N$ is discovered. For example, when $p = 5$, $n = 7$, $k = 4$, and $x = 4$, the sequence is

4, 569, 129, 444, 254, 319, 379, 194, 504, 69

the period of which is 10.

Since determining the period is a fundamental problem for engineering applications such as cryptography and pseudorandom numbers, it is important to know the period instead of calculating $x$, $T_n(x) \mod p^k$, $T_n^2(x) \mod p^k$, $\cdots$, until $T_n^N(x) \equiv x \mod p^k$, which requires at most $p^k$ times calculations of $T_n(x) \mod p^k$. We study this in the following section.

### 2.2. The period of Chebyshev polynomial sequences modulo $p^k$

When $R = \mathbb{Z}/2^k\mathbb{Z}$, we have already shown an interesting property of Chebyshev polynomial sequences: the period of $\{T_n(x) \mod 2^{k-1}\}_{i=0}^{N-1}$ is twice as long as that of $\{T_n(x) \mod 2^{k-2}\}_{i=0}^{N-1}$ [9]. Therefore, it is expected that the period of $\{T_n(x) \mod p^{k-1}\}_{i=0}^{N-1}$ is $p$ times as long as that of $\{T_n(x) \mod p^{k-2}\}_{i=0}^{N-1}$ for any odd prime $p$. However, this is not always true unlike the case of even prime, which is shown by numerical examples. The periods $N$ of $T_n(x)$ mod $p^k$ for $x = 2, 3, 4, 5, 6$ are shown in Tables 1, 2, and 3, where $(p, n) = (3, 5), (5, 7)$ and $(7, 5)$, respectively. For example, such periodic property does not hold when $x = 3, 6$ and $k = 2$ in the case of $(p, n) = (3, 5)$. Meanwhile, it can be seen from these Tables that the period of sequences modulo $p^k$ is $p$ times as long as that in the operation of modulo $p^{k-1}$ when $k \geq 4$. In this section, we seek conditions for such periodic properties of Chebyshev polynomial sequences modulo $p^k$.

Assume $X$ and $w \geq 1$ satisfy the relation

$$T_n(X) \equiv X \mod p^w,$$  \hspace{1cm} (8)

$$T_n(X) \not\equiv X \mod p^{w+1}.$$  \hspace{1cm} (9)

According to (8), there exits an integer $b \in \{1, 2, \cdots, p-1\}$ such that

$$T_n(X) = X + b \cdot p^w.$$  \hspace{1cm} (9)

Let $\mathcal{G}$ be a finite field with characteristic $p$. The order of an element $m$ in the group $\mathcal{G}$, denoted as $\text{ord}(m)$, is the least positive number such that $m^{\text{ord}(m)} \equiv 1 \mod p$. These lemmas will be used in the following discussion.

**Lemma 2** Let $p$ be a prime number. For any integer $x \not\equiv 0, 1$,

$$1 + x + x^2 + \cdots + x^{\text{ord}(x)-1} \mod p \equiv 0$$  \hspace{1cm} (10)

**Proof:** Since $x^{\text{ord}(x)} - 1 = (x-1)(x^{\text{ord}(x)-1} + x^{\text{ord}(x)-2} + \cdots + x + 1) \equiv 0 \mod p$ and $x - 1 \not\equiv 0 \mod p$, (10) holds. \hfill $\Box$

**Lemma 3** Let $p$ be a prime number. For any integer $x \not\equiv 0, 1$,

$$1 + x + x^2 + \cdots + x^{p-1} \mod p \equiv \begin{cases} 1, & x \not\equiv 1 \\ 0, & x = 1 \end{cases}$$  \hspace{1cm} (11)
Table 1: The list of periods $N$ for several values of $x$ and $k$, where $p = 3$ and $n = 5$.

<table>
<thead>
<tr>
<th>$x = 2$</th>
<th>$x = 3$</th>
<th>$x = 4$</th>
<th>$x = 5$</th>
<th>$x = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>9</td>
<td>18</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$k = 5$</td>
<td>27</td>
<td>54</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

Proof: Using $x^p - 1 = (x - 1)(x^{p-1} + x^{p-2} + \cdots + x + 1)$ together with Fermat’s little theorem, $x^n \equiv x \mod p$, it is easy to see that the equation (11) holds. □

Let $T_n'(x)$ be the derivative of $T_n(x)$ with respect to $x$.

**Lemma 4** For any $x$, $T_n'(x) \equiv 0 \mod p$.

**Proof:** We just recall the result of [12]. If there is an integer $x$ such that $f'(x) \equiv 0 \mod p$, the number of solutions of $f(x) \equiv 0 \mod p$ is none, where $f(x)$ is any integral polynomial. Since $T_n(x) \mod p$ is a permutation polynomial, there is an integer such that $T_n(x) \equiv 0 \mod p^k$. Thus, the assertion is verified. □

Firstly, we show the following lemma, which means that the period of $T_n(X) \mod p^{x+1}$ is related to the value of $T_n'(X) \mod p$.

**Lemma 5** Assume $X$ and $w \geq 1$ satisfy the relation of (8). Let $\ell = \text{ord}(T_n'(X))$ for $T_n'(X) \not\equiv 1 \mod p$ and $\ell = \ell$, otherwise. Then, $T_n'(X) \equiv X \mod p^{x+1}$.

**Proof:** Substituting (9) into $T_n(x) = a_1x + a_3x^3 + \cdots + a_nx^n$ gives

$$T_n'(X) = a_1(X + b \cdot p^w) + a_3(X + b \cdot p^w)^3 + \cdots + a_n(X + b \cdot p^w)^n$$

$$\equiv T_n(x) + b \cdot p^w \cdot T_n'(X) \mod p^{x+1}$$

$$\equiv X + b \cdot p^w \cdot (T_n'(X) + 1) \mod p^{x+1}.$$ (12)

Repeating the above, it holds that

$$T_n'(X) \equiv X + b \cdot p^w \sum_{m=0}^{i-1} T_n(X)^m \mod p^{x+1}.$$ (13)

where integer $i \geq 1$. By virtue of Lemmas 2 and 3, the assertion is verified. □

Let us define $G$ as $G(x) = T_n'(x) = T_n(x)$ for a positive integer $i \geq 1$. From the semi-group property of Chebyshev polynomials, $G$ is also a Chebyshev polynomial of odd degree. The chain rule is a formula for computing the derivative of the composition of two functions $f$ and $g$, that is, $(f(g(x)))' = f'(g(x)) \cdot g'(x)$.

**Lemma 6** Assume $X$ and $w \geq 1$ satisfy the relation of (8). Let $G = T_n$. For $i \geq 1$,

$$G'(X) \equiv (T_n'(X))^i \mod p.$$ (14)

Proof: Using mathematical induction together with $T_n(X) \equiv X \mod p$ leads to (14). □

When $w \geq 2$, (13) also holds in the operation of modulo $p^{x+2}$. Thus, we have the following lemma.

**Lemma 7** Assume $X$ and $w \geq 2$ satisfy the relation of (8). If $T_n'(X) \equiv 1 \mod p$, then $T_n'(X) \equiv X \mod p^{x+1}$ and $T_n'(X) \not\equiv X \mod p^{x+2}$.

Next, we show a condition of $X$ for which the period of sequence $(T_n'(X) \mod p^{x+1})$ becomes $p$ times longer as $k$ increases.

**Lemma 8** Let $m$ be a positive integer. Assume $X$ and $w \geq 2$ satisfy the relation of (8). If $T_n'(X) \equiv 1 \mod p$, then

$$T_n^{pm}(X) \equiv X \mod p^{x+m},$$

$$T_n^{pm}(X) \not\equiv X \mod p^{x+m+1}.$$ (15)

**Proof:** We prove the above lemma by mathematical induction. It is shown by Lemma 7 that the case of $m = 1$ is satisfied. Suppose that (15) is true when $m = s$. From the semi-group property of $T_n$, $G = T_n'$ is also a Chebyshev polynomial. Using (14), $G'(X) \equiv 1 \mod p$. Therefore, by Lemma 7, we have $G^{p}(X) \equiv X \mod p^{x+2}$ and $G^{p}(X) \not\equiv X \mod p^{x+3}$, which means (15) is also true with $m = s + 1$ since $G^{p} = T_n^{p+1}$. From the above discussions, (15) is satisfied for arbitrary $m \geq 1$, which completes the proof. □

As a direct consequence of the above lemma, the period of Chebyshev polynomial sequences modulo $p^k$ is derived as $N = p^{k+2}$ under the assumption that the condition of (8) is satisfied.

**Example 1** When $p = 3$, $n = 5$ and $X = 2$, we obtain $w = 2$ since $T_2(2) \equiv 2 \mod 3$ and $T_3(2) \equiv 2 \mod 3$. Since $T_2'(X) \equiv 1 \mod 3$, the period $N$ of the sequence $(T_2(X) \mod 3^k)$ is derived as $N = 3^{k+2}$ for $k \geq 2$. 

Table 2: The list of periods $N$ for several values of $x$ and $k$, where $p = 5$ and $n = 7$.

<table>
<thead>
<tr>
<th>$x = 2$</th>
<th>$x = 3$</th>
<th>$x = 4$</th>
<th>$x = 5$</th>
<th>$x = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>20</td>
<td>20</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>$k = 5$</td>
<td>100</td>
<td>100</td>
<td>50</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 3: The list of periods $N$ for several values of $x$ and $k$, where $p = 5$ and $n = 7$.

<table>
<thead>
<tr>
<th>$x = 2$</th>
<th>$x = 3$</th>
<th>$x = 4$</th>
<th>$x = 5$</th>
<th>$x = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 1$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$k = 2$</td>
<td>6</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$k = 3$</td>
<td>3</td>
<td>42</td>
<td>21</td>
<td>6</td>
</tr>
<tr>
<td>$k = 4$</td>
<td>294</td>
<td>147</td>
<td>42</td>
<td>147</td>
</tr>
<tr>
<td>$k = 5$</td>
<td>2058</td>
<td>1029</td>
<td>1029</td>
<td>294</td>
</tr>
</tbody>
</table>
Finally, we have the following theorem.

**Theorem 2** Assume $X$ and $w$ satisfy the relation of (8). Let $\ell = \text{ord}(T_n(X))$ for $T_n(X) \not\equiv 1 \mod p$ and $\ell = p$, otherwise. Then, there is an integer $w_2$ such that $T_n(X) \equiv X \mod p^{w_2}$ and $T_n(X) \not\equiv X \mod p^{w_2+1}$. Furthermore, the period of Chebyshev polynomial sequences $\{T_n(X)\mod p^{i-1}\}$ is derived as $N = \ell \cdot p^{w_2}$ for $k \geq w_2$.

When $T_n(X) \not\equiv X \mod p$, there is a least positive integer $s$ such that $T_n(X) \equiv X \mod p^s$. Let $G = T_n$, then there is an integer $w_2$ such that $G(X) \equiv X \mod p^{w_2}$ and $G(W) \not\equiv X \mod p^{w_2+1}$. Since $\ell$ is the order of the period $n \mod p$, $G(X) = (T_n(X))^\ell \equiv 1 \mod p$. Together with Lemma 8, $G(X) \equiv (X) \mod p\ell$, which implies that the period of Chebyshev polynomial sequence $\{T_n(X)\mod p^{i-1}\}$ must be $N = \ell \cdot p^{w_2}$.

**Proof:** First, we consider the case for $X$ with $w \geq 1$. Assume $T_n(X) \equiv X \mod p^{w+1}$. Let $G = T_n$, then, there is an integer $w_2 \geq 2$ such that $G(X) \equiv X \mod p^{w_2}$ and $G(W) \not\equiv X \mod p^{w_2+1}$. Since $\ell$ is the order of the period $n \mod p$, $G(X) = (T_n(X))^\ell \equiv 1 \mod p$. Together with Lemma 8, $G(X) \equiv (X) \mod p\ell$, which implies that the period of Chebyshev polynomial sequence $\{T_n(X)\mod p^{i-1}\}$ must be derived for any $k \geq w_2$.

For the case that $T_n(X) \not\equiv X \mod p$, $G(X) \equiv X \mod p$. By the same argument as the above, the period of sequence $X, G(X) \equiv (X)^\ell \equiv X \mod p^{w_2}$. Thus, we obtain $N = s \cdot \ell \cdot p^{w_2}$ for $k \geq w_2$.

Using Theorem 2, forming a sequence as $X, T_n(X) \mod p^2, \ldots \mod p^N(X)$ is not needed to investigate the period. We just try to find the value of $w_2$ from $w_2 = 1$ to $k$ and the order of derivatives $T_n(X) \mod p^2$. After finding $w_2$, the period of sequence $\{T_n(X)\mod p^{i-1}\}$ is derived for any $k \geq w_2$.

**Example 2** When $p = 5, n = 7$ and $X = 2$, we obtain $w = 1$. Since $T_2(X) \equiv 2 \mod 2$, $T_2(X) \equiv 1 \mod 5$, $\ell = 4$. Since $T_2(X) \equiv 2 \mod 2$ and $T_2(X) \equiv 2 \mod 2$, the period $N$ is derived as $N = 4 \cdot 5^4$ for $k \geq 3$.

**Example 3** When $p = 7, n = 5$, and $X = 2$, $T_2(X) \not\equiv X \mod p$. Define $G = T_2$, then $G(2) \equiv 2 \mod 7$ and $G(2) \equiv 4 \mod 7$. Since $4^2 \equiv 7 \equiv 1$, $\ell = 3$ and $G(2) \equiv 2 \mod 7^2$. Therefore, the period $N$ of sequence $\{T_2(X)\mod p^{i-1}\}$ is derived as $N = 2 \cdot 3 \cdot 7^2$ for $k \geq 2$.

One can also see that the numerical results in Tables 1, 2, and 3 are consistent with our theoretical results.

3. Conclusion

In this paper, we showed that Chebyshev polynomials become permutation polynomials over the residue rings of integers of powers of odd prime. We also derive the periodic property of Chebyshev polynomial sequences over the ring under some conditions. The result is useful for finding the period of Chebyshev polynomial sequences.

The detailed analysis of the relation of the period, degree, and initial value of Chebyshev polynomials over powers of odd prime is much more complicated. This topic is challenging and needed further research.

References


Pseudo Random Binary Sequence Generated by Trace and Legendre Symbol with Non Primitive Element in $\mathbb{F}_p$^2

C. Ogawa, A. M. Arshad, Y. Nogami, S. Uehara, K. Tsuchiya, and R. M. Zaragoza

† Graduate School of Natural Science and Technology, Okayama University, 3-1-1, Tsushima-naka, Kita, Okayama, 700-8530, E-mail: yasuyuki.nogami@okayama-u.ac.jp.
‡ Graduate School of Environmental Engineering, The University of Kitakyushu, 1-1, HibiKino, Wakamatsu, Kitakyushu, Fukuoka, 808-0135, E-mail: uehara@kitakyu-u.ac.jp.
‡‡ Koden Electronics Co. Ltd., 2-13-24 Tamagawa, Ota-ku, Tokyo, 146-0095, E-mail:k-tsuchiya@koden-electronics.co.jp.

Abstract—Pseudo binary random sequence has many uses such as nonce for security applications. Some of them needs to have long period and high linear complexity. The authors have proposed a generation method that uses primitive polynomial, trace function, and Legendre symbol over odd characteristic field. The preparation of primitive polynomial is not always easy. This paper shows that some non-primitive irreducible polynomials generate the same random binary sequence generated by a certain primitive polynomial. Then, some example are also introduced.

1. Introduction

There are many kinds of pseudo binary random sequence generated over finite fields. Among them, maximal length sequence (M-sequence) and Legendre sequence are well known [1],[2]. M-sequence uses trace function and Legendre sequence uses Legendre symbol. Their typical properties such as period, autocorrelation, and linear complexity have been theoretically shown. The authors have proposed a pseudo binary random sequence generated by primitive polynomial, trace function, and Legendre symbol [3]. It has long period and high linear complexity. These properties have been theoretically shown. Different from M-sequence and Legendre sequence, this sequence has two parameters $p$ and $m$, where $p$ and $m$ are the characteristic and extension degree by which the base extension field $\mathbb{F}_p$^m is defined. In addition, in the same of M-sequence, it also needs a primitive polynomial.

In order to prepare a long period sequence for some cryptographic applications, the characteristic $p$ or the extension degree $m$ should be large. Accordingly, the previous sequence needs to prepare a primitive polynomial of degree $m$ over $\mathbb{F}_p$. However, the preparation is not always easy. This paper shows some non-primitive irreducible polynomials are able to generate the same sequence generated by a certain primitive polynomial. If the condition is clearly given, the preparation of the non-primitive irreducible polynomial will be easier than that of primitive polynomial. When the degree $m$ is restricted to 2, this paper not only considers the conditions but also shows some examples.

2. Preparation

This section briefly introduces some mathematical tools. Throughout this paper, $p$ be an odd prime number.

2.1. Irreducible and primitive polynomials

Let $\mathbb{F}_p$ be a prime field of odd characteristic $p$. When $f(x)$ of degree $m$ over $\mathbb{F}_p$ is not factorized into smaller degree polynomials over $\mathbb{F}_p$, it is called irreducible polynomial. Let $\omega$ be its zero, $\omega$ belongs to the extension field $\mathbb{F}_p^m$ and its order $e$ is a divisor of $p^m - 1$. It is noted that $p^m - 1$ is the order of the multiplicative group $\mathbb{F}_p^*$, $\mathbb{F}_p^* = \mathbb{F}_p - \{0\}$. Particularly when $e = p^m - 1$, it is called a primitive polynomial and its zero is called a primitive element in $\mathbb{F}_p^m$ correspondingly. M-sequence and our previous work [3] utilize a primitive element to generate a maximal length sequence because the primitive element $\omega$ is able to represent all non-zero elements as its power $\omega^i$, $i = 0, 1, 2, \cdots, p^m - 2$. When $m = 2$, an irreducible polynomial of degree 2 over $\mathbb{F}_p$ is easily generated even if $p$ is large.

2.2. Trace function and Legendre symbol

Consider an extension field $\mathbb{F}_p^m$. Then, trace function for $X \in \mathbb{F}_p^m$ is defined as follows.

$$x = \text{Tr}(X) = \sum_{i=0}^{m-1} X^i, \quad (1)$$

$x$ becomes an element in $\mathbb{F}_p$ and the above trace function has a linearity over $\mathbb{F}_p$ as follows.

$$\text{Tr}(aX + bY) = a\text{Tr}(X) + b\text{Tr}(Y), \quad (2)$$

where $a, b \in \mathbb{F}_p$ and $Y \in \mathbb{F}_p^m$. In the previous work [3], trace function is used for mapping a vector in $\mathbb{F}_p^m$ to a scalar
in $\mathbb{F}_p$. Then, Legendre symbol is calculated as follows.

$$\left( \frac{a}{p} \right) = a^{(p-1)/2} \mod p$$

$$= \begin{cases} 0 & \text{when } a = 0, \\ 1 & \text{if } a \text{ is a non-zero QR}, \\ -1 & \text{otherwise, that is } a \text{ is a QNR}, \\ \end{cases}$$  

(3)

where QR and QNR are abbreviations of quadratic residue and quadratic non-residue, respectively. In our previous work, Legendre symbol is used for mapping a scalar in $\mathbb{F}_p$ to a signed binary value such as $[0, 1, -1]$.

2.3. Previous work

The previous work [3] has proposed a pseudo random binary sequence generated by using primitive polynomial, trace function, and Legendre symbol as follows.

$$\mathcal{T} = \{t_i; t_i = f \left( \left( \text{Tr} \left( \omega^i \right) \right)_p \right) \}; \quad i = 0, 1, 2, \cdots .$$  

(4)

where $f(\cdot)$ is defined as

$$f(x) = \begin{cases} 0 & \text{if } x = 0, 1, \\ 1 & \text{otherwise}. \\ \end{cases}$$  

(5)

$\omega$ in Eq. (4) is a primitive element in $\mathbb{F}_{p^e}$. Then, its period is given by $2(p^m - 1)/(p - 1)$.

Let the autocorrelation with shift value $x$ be defined by

$$R_T(x) = \sum_{i=0}^{p-1} (-1)^{i+x-i},$$  

(6)

the autocorrelation of $\mathcal{T}$ is given by

$$R_T(x) = \begin{cases} \frac{2(p^m - 1)}{p - 1} & \text{if } x = 0, \\ -2p^{m-1} + \frac{2(p^{m-1} - 1)}{p - 1} & \text{else if } x = n/2, \\ \frac{2(p^{m-2} - 1)}{p - 1} & \text{otherwise}. \\ \end{cases}$$  

(7)

As a small example, Figure. 1 shows the graph of the autocorrelation $R_T(x)$ with $p = 7$ and $m = 2$.

3. Binary sequence with non-primitive polynomial

This paper introduces, particularly when the extension degree $m = 2$, some non-primitive irreducible polynomials generate the same random binary sequence generated by a certain primitive polynomial.

3.1. Motivation

First of all, when the characteristic $p$ or the degree $m$ are large such as used for cryptographies, preparing a primitive polynomial is not always easy. Let us consider the case that $p$ is a large prime number and $m = 2$. In this case, consider all prime factors $p_i$ of $p^2 - 1$ as

$$p^2 - 1 = \prod_i p_i^{e_i},$$  

(8)

then check the following relation for every $p_i$.

$$f(x) \mid x^{(p^2-1)/p_i} - 1,$$  

(9)

where $f(x)$ is a randomly generated irreducible polynomial of degree 2 over $\mathbb{F}_p$.

On the other hand, generating an irreducible polynomial of an arbitrary degree over $\mathbb{F}_p$ is not difficult [4]. Particularly, when every factor of the degree $m$ divides $p - 1$, it becomes quite easy. When $m = 2$ as an example, using $c \in \mathbb{F}_p$ such that $(c/p) = -1$, where

$$f(x) = x^2 - c$$  

(10)

becomes an irreducible polynomial over $\mathbb{F}_p$. Thus, it is more practical that the same binary sequence is generated by using some non-primitive irreducible polynomial.

3.2. Example

Let us observe a small example with $p = 7$ and $m = 2$. Table 1 shows the result. As introduced in the previous section, irreducible binomials such as $x^3 - 3$, $x^2 - 5$, and $x^2 - 6$ are obtained. Applying a simple substitution such as $x \rightarrow x+1$, irreducible trinomials such as $x^2 + 2x + 5$ are obtained. Among them, there are primitive or non-primitive irreducible polynomials as shown in Table 1.

See the row (1) of the table. In this case, $x^2 + 2x + 5$, $x^2 + 4x + 6$, and $x^2 + x + 3$ are transformed from $x^2 - 2$ and generate the same binary sequence 0100011101101001. Among these three irreducible polynomials, $x^2 + 2x + 5$ and $x^2 + x + 3$ are primitive polynomials of order $e = 7^2 - 1 = 48$. On the other hand, $x^2 + 4x + 6$ is a non-primitive polynomial of order 16, however it generates the same binary sequence.
Table 1: Binary sequence generated by primitive polynomial and irreducible polynomial with $p = 7$ and $m = 2$

<table>
<thead>
<tr>
<th></th>
<th>$x^3 - 3$</th>
<th>$x^2 - 5$</th>
<th>$x^2 - 6$</th>
<th>$x^{24}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>$x \leftarrow x + 1$</td>
<td>$x^2 + 2x + 5$</td>
<td>$x \leftarrow x + 2$</td>
<td>$x \leftarrow x + x + 3$</td>
</tr>
<tr>
<td>(2)</td>
<td>$x \leftarrow x + 6$</td>
<td>$x^2 + 5x + 5$</td>
<td>$x \leftarrow x + 5$</td>
<td>$x^2 + 3x + 6$</td>
</tr>
<tr>
<td>(3)</td>
<td>$x \leftarrow x + 3$</td>
<td>$x^2 + 6x + 6$</td>
<td>$x \leftarrow x + 6$</td>
<td>$x^2 + 5x + 3$</td>
</tr>
<tr>
<td>(4)</td>
<td>$x \leftarrow x + 4$</td>
<td>$x^2 + x + 6$</td>
<td>$x \leftarrow x + 1$</td>
<td>$x^2 + 2x + 3$</td>
</tr>
</tbody>
</table>

(*) They are non-primitive irreducible polynomials over $F_7$. The others are all primitive polynomials.

3.3. Consideration

Since Table 1 is a small example, the primitivity of irreducible polynomial could be easily checked. However, when the characteristic $p$ is large, the primitivity check is not always easy. According to Table 1, it is found that an irreducible polynomial of order 16 generates the same binary sequence generated by a certain primitive polynomial. In detail, it has been found that the non-primitive irreducible polynomials marked with (*) in Table 1 have the same order 16. The authors have tested a lot of prime numbers as the characteristic $p$ with extension degree $m = 2$. According to the results, without any counter examples, the orders of the non-primitive polynomials have been given by $(p^2 - 1)/s$ and $s$ is an odd prime factor of $p^2 - 1$.

4. Conclusion and future works

This paper has shown that, when the degree is restricted to 2, some non-primitive irreducible polynomials are able to generate the same binary sequence generated by a certain primitive polynomial. It means that, if the condition for the non-primitive irreducible polynomials are shown clearly, primitive polynomials are not necessary for generating maximal length sequence. As a future work, the condition should be theoretically shown.

Acknowledgments

This work has been supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (A) Number 16H01723.

References

Performance Analysis of the Interval Algorithm for Random Number Generation in the Case of Markov Coin Tossings

Yasutada Oohama†

†University of Electro-Communications, 1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585, Japan
Email: oohama@uec.ac.jp

Abstract—In this paper we analyze the interval algorithm for random number generation proposed by Han and Hoshi in the case of Markov coin tossings. Using the expression of real numbers on the interval [0,1), we first establish an explicit representation of the interval algorithm with the representation of real numbers on the interval [0,1) based on one number systems. Next, using the expression of the interval algorithm, we give a rigorous analysis of the interval algorithm. We discuss the difference between the expected number of the coin tosses in the interval algorithm and their upper bound derived by Han and Hoshi and show that it can be characterized explicitly with the established expression of the interval algorithm.

1. Introduction

Simulation problems of generating random sequences from a prescribed information source by using a random sequence from a given information source are called the random number generation. In the random number generation random sequences from a prescribed information sources are called the target random sequences which we wish to produce and the random sequence from given information sources are called the coin random sequences that the target random sequences are made from.

There have been several works on the random number generation in the field of computer science and information theory. Some interesting relations between random number generation and information theory have been found in the papers of Elias [1] and Knuth and Yao [2].

Han and Hoshi [3] studied a variable-to-fixed random number generation problem. They studied the method of generating target random sequences of fixed length from a prescribed information source by using coin random sequences of variable length from a given information source. They proposed a simple algorithm called the interval algorithm and obtained results for its performance analysis. They established an upper bound of the average length of coin random sequences necessary to create target random sequences. The derived bound is characterized with a fraction of two entropies of given and prescribed sources and is shown to be asymptotically optimal for large length of output sequences.

In our previous work [4], we studied the performance analysis of the interval algorithm for random number generation proposed by Han and Hoshi [3]. In this work we treated the case where we wish generate a target random variable by using a coin random sequence from a stationary memoryless sources. In this paper we study the performance analysis of the interval algorithm in an extended case where coin random sequences are from the stationary Markov information sources.

In [4], we derived explicit results on the performance analysis of the interval algorithm for random number generation using an expression of real numbers in the unit interval [0,1). On the expression of real numbers in the unit interval, we used a kind of generalized number system based on the stochastic structure of the coin random process. Using the above representation of real numbers on the interval, we established an explicit expression of the interval algorithm. In this paper we show that the same result also holds for the case where the coin random process is a Markov chain. Using this expression of the algorithm, we give a rigorous analysis of the interval algorithm. We discuss the difference between the expected number of the coin tosses in the interval algorithm and their upper bound derived by Han and Hoshi and show that it can be characterized explicitly with the established expression of the interval algorithm.

2. Interval Algorithm for Random Number Generation

Let $X$ be random variables taking values in a finite set $X = \{0, 1, \ldots, N - 1\}$. Let $p_\times \triangleq \{p_\times(x)\}_{x \in X}$ be a probability distribution of $X$. Let $\{Y_t\}_{t\geq0}$ be a stationary Markov source. For each $t = 1, 2, \ldots$, $Y_t$ takes values in a finite set $Y = \{0, 1, \ldots, M - 1\}$. The stationary Markov source $\{Y_t\}_{t\geq0}$ is specified with the $M \times M$ stochastic matrix denoted by $P = [P_{ij}]$, where $P_{ij} = \Pr(Y_{t+1} = j|Y_t = i)$, for $t = 1, 2, \ldots$.

We also write $P_{ij}, (i, j) \in Y^2$ as $P_{ij} = p_{ij}$. Let $Y^\ast$ denote the set of all finite sequence emitted from the above information source. We write a string from information source as $y^m = y_{t+1} \cdots y_m \in Y^\ast$. If $l > m$, the string $y^m$ means null string denoted by $\Lambda$. When $l = 1$, we frequently omit...
the suffix 1 of $y_m$ and write $y^m = y_1 y_2 \cdots y_m$. Let $p_Y(y^m)$ denote the probability of $y^m$. Since the information source is a stationary Markov source, we have

$$p_Y(y^m) = p_Y(y_1) P_{Y|Y_{1}} Y_2 \cdots P_{Y|Y_{1} \cdots Y_{m-1}} Y_m.$$

Here $[p_Y(\alpha)]_{\alpha \in \mathcal{Y}}$ is a stationary distribution computed from $P$. The probability of the null string $\lambda$ assumes to be one.

In this paper we deal with the variable to fixed random number generation problem of generating target random variable $X$ by using the coin random sequence $Y_1 Y_2 \cdots Y_i \cdots$ from a stationary Markov information sources $\{Y_i\}_{i=1}^{\infty}$. A formal definition of the variable to fixed random number generation problem is the following. Repeated tosses of the coin random variable $Y$ produces random sequence $Y_1 Y_2 \cdots$ from a discrete memoryless source. The coin toss terminates at some finite time $L$ to generate a random variable $X$ with a prescribed distribution $p_X$. $L$ is a random variable specified in terms of a deterministic two valued function such that $f(Y) = \text{‘Continue’}$ for $1 \leq i \leq L - 1$ and $f(Z) = \text{‘Stop’}$. The output $X$ is expressed as $X = \psi(Y^L)$ with some deterministic function $\psi$.

In the above random number generation problem Han and Hoshi [3] proposed a simple algorithm called interval algorithm and evaluated its performance. Let $I = [0, 1)$. Define the cumulative probabilities for $p_Y$ by

$$c_Y(0) \overset{\Delta}{=} 0, c_Y(y) \overset{\Delta}{=} \sum_{y \leq t \leq y} p_Y(t), 1 \leq y \leq M - 1.$$  

Using these probabilities, define the decomposition of $I$ by

$$I_Y(y) \overset{\Delta}{=} [c_Y(y), c_Y(y) + p_Y(y)).$$

For $p_X$, we use the same notations and definitions as those for $p_Y$. For given $y_1 \in \mathcal{Y}$, define the cumulative probabilities for $p_Y([y_1) = \{p_Y(y_1 y_2)\}_{y_2 \in \mathcal{Y}}$ by

$$c_Y(0)_{y_1} \overset{\Delta}{=} 0, c_Y(y_2\mid y_1) \overset{\Delta}{=} \sum_{y_1 \leq t \leq y_2} p_Y(t\mid y_1), 1 \leq y_2 \leq M - 1.$$  

For $k = 1, 2, \cdots$, and any string $y^k = y_1 y_2 \cdots y_k \in \mathcal{Y}^k$, define the semi-open interval $I_Y(y^k) \overset{\Delta}{=} [L_Y(y^k), U_Y(y^k))$ by the following recursions:

$$
\begin{align*}
L_Y(y_1) &= c_Y(y_1), \\
U_Y(y_1) &= c_Y(y_1) + p_Y(y_1) \\
L_Y(y') &= L_Y(y'+1) + p_Y(y'+1) c_Y(y|y'+1), \\
U_Y(y') &= L_Y(y') + p_Y(y'), \quad \text{for } 2 \leq i \leq k.
\end{align*}
$$

The procedure of computing upper and lower end points of the interval corresponding to a given sequence is equivalent to the encoding algorithm in the arithmetic coding.

Interval algorithm by Han and Hoshi [3] can be stated in the following.

Interval Algorithm (Han and Hoshi [3]):

1) Set $i = k = 1, y_0 = \lambda$.

2) Given $y_{i-1}$, generate a letter $y_i \in \mathcal{Y}$ according to the transition probability $P_{Y|Y_{i-1}}$ of the coin random variable.

3) Compute $I_Y(y') = [L_Y(y'), U_Y(y')]$ according to the recursion (1).

4) If $I_Y(y') \subseteq I_X(x)$ for some $x \in \mathcal{X}$, then output $x$ as the value of target random variable $X$ and stop the algorithm.

5) Set $i = i + 1$ and go to 2).

In the above interval algorithm the target random variable $X$ can exactly be produced.

3. An Explicit Representation of the Interval Algorithm

In this section we give two expressions of real numbers in the interval $I = [0, 1)$ on the number system. There is some complementary relation between the above two expressions. Using those expressions we give an explicit form of the interval algorithm.

3.1. Representation of real numbers

For $z \in [0, 1)$, define the sequence $\{a_i\}_{i=1}^{\infty} \in \mathcal{Y}^*$ such that

$$z \in I_Y(a_i), i = 1, 2, \cdots.$$  

It can easily be verified that using $a_1, a_2, \cdots, z$ can be expressed in the following manner:

$$z = \sum_{k=1}^{\infty} p_Y(a^{k-1}) \sum_{a \prec a_k} p_Y(a_{k-1}) = \sum_{k=1}^{\infty} p_Y(a^{k-1}) c_Y(a_k|a_{k-1}).$$

We call the above expression the $p_Y$-ary representation of the real number $z$ and write as

$$z = 0. a_1 a_2 a_3 \cdots .$$

In the above expression, if we wish to express $z$ with the sum of the number having the expression

$$0. a_1 a_2 a_3 \cdots a_0 0 \cdots$$

and the other remaining term, we write

$$z = 0. a_1 a_2 a_3 \cdots a_t + 0. a_{t+1} a_{t+2} \cdots$$

where the second term is defined by

$$0. a_{t+1} a_{t+2} \cdots a_{k-1} \overset{\Delta}{=} \sum_{k=t+1}^{\infty} p_Y(a^{k-1}) c_Y(a_k|a_{k-1}).$$

Next, for $z \in [0, 1)$, set $\tilde{z} = 1 - z$. Using the sequence $\{a_i\}_{i=1}^{\infty}$ appearing in the $p_Y$-ary representation of the real number $z$, $\tilde{z}$ has an expression

$$\tilde{z} = \sum_{k=1}^{\infty} p_Y(a^{k-1}) \sum_{a \succ a_k} p_Y(a_{k-1}).$$
Then, adopting the notation
\[ c_Y(\bar{a}|a_{k-1}) \triangleq \sum_{i>a} p_Y(i|a_{k-1}), \]
we obtain the following expression
\[ \bar{z} = \sum_{k \geq 1} p_Y(a^{k-1}c_Y(\bar{a}|a_{k-1})). \]

We call the above expression the \textit{p}_Y\text{-ary co-representation} of the real number \( \bar{z} \) and write as
\[ \bar{z} = 0.\bar{a}_1\bar{a}_2\cdots \]

Let \( z^{(n)} \) denote the real number which is obtained by rounding \( \bar{z} \) to \( n \)-digits in the \( p_Y\)-ary representation, that is,
\[ z^{(n)} \triangleq 0. a_1 a_2 \cdots a_n. \]

Similarly, let \( \bar{z}^{(n)} \) denote the real number which is obtained by rounding off \( \bar{z} \) to \( n \)-digits in the \( p_Y\)-ary co-representation, that is,
\[ \bar{z}^{(n)} \triangleq 0. \bar{a}_1 \bar{a}_2 \cdots \bar{a}_n. \]

It can easily be verified that the \( p_Y\)-ary representation and the \( p_Y\)-ary co-representation of the real number \( z \) satisfy the following.

\textbf{Property 1}

a) For any \( i, z \in I_Y(a^i) \).

b) \[ c_Y(a_i|a_{i-1}) + p_Y(a_i|a_{i-1}) = 1 - p_Y(a_i|a_{i-1}). \]

c) For \( z = 0. a_1 a_2 \cdots a_n \in [0, 1) \), we have
\[ z^{(n)} + \bar{z}^{(n)} = 1 - p_Y(a^i). \]

\subsection*{3.2. An explicit representation of the interval algorithm}

In this subsection, we give an explicit form of the interval algorithm by using the \( p_Y\)-ary representation and \( p_Y\)-ary co-representation of the real number in the interval \( I = [0, 1) \). It can easily be seen from the definition of the interval algorithm the interval \( I_x(x) = [L_x(x), U_x(x)] \) corresponding to the target random variable \( X \) takes values in \( X \). Suppose that lower and upper endpoints \( L_x(x) \) and \( U_x(x) \) have the following \( p_Y\)-ary representation and \( p_Y\)-ary co-representations:

\[ L_x(x) = 0. a_1 a_2 \cdots, \quad \bar{L}_x(x) = 0. \bar{a}_1 \bar{a}_2 \cdots, \]
\[ U_x(x) = 0. b_1 b_2 \cdots. \]

For each \( x \in X \), there exists an integer \( t = t(x) \) such that representations of \( L_x(x) \) and \( U_x(x) \) have different values at the \( t \)-th place at their \( p_Y\)-ary representations. Then, we have
\[ p_x = p_Y(a^{t-1}d_Y(a_t|a_{t-1}) \]
\[ + \sum_{k \geq t+1} \left( p_Y(a^{k-1}c_Y(\bar{a}_k|a_{k-1}) + p_Y(b^{k-1}_t c_Y(b_t|b_{t-1})) \right), \]

where
\[ d_Y(a_t, b_t|a_{t-1}) \triangleq \sum_{a_i < a_t} p_Y(a_i|a_{i-1}) \]
and \( d_Y(a_t, b_t|a_{t-1}) = 0 \) when \( b_t = a_t + 1 \). The above expression leads to the following description of \( I_x(x) \) with the disjoint sum of intervals corresponding to the target random sequences in the interval algorithm:

\[ I_x(x) = \sum_{a_i < c_h} I_y(a^{i-1}y) \]
\[ + \sum_{k \geq t+1} \left( \sum_{y < a_h} I_x(a^{k-1}y) + \sum_{y < b_h} I_y(b^{k-1}y) \right). \]

It can be seen from the above presentation that the interval \( \sum_{a_i < c_h} I_y(a^{i-1}y) \) is in the middle of the interval \( I_y(x) \) and that the sequence of intervals \( \left\{ \sum_{y < a_h} I_x(a^{i-1}y) \right\}_{k \geq t} \) entirely covers the lower part of the interval \( I_y(x) \). These intervals are called \textit{downward sequences} in Han and Hoshi [3]. We also know that the sequence of intervals \( \left\{ \sum_{y < c_h} I_y(b^{k-1}y) \right\}_{k \geq t+1} \) in the third term in the right member of the above equation entirely covers the upper part of \( I_y(x) \). This sequence of the intervals are called \textit{upward sequence} in Han and Hoshi [3]. The result of Theorem can be regarded as giving an explicit form of upward/downward sequences of intervals in the interval algorithm. Those sequences of intervals is shown in Fig. 1.

As a corollary of this theorem we can obtain a result, which is quite useful for the performance algorithm of the interval algorithm. To describe this result we define some quantities: For each \( a \in \{1, 2, \cdots, M - 1\} \), let \( \{l_{a,k} \}_{k \geq 1} \) be a sequence of positive integers satisfying

\[ t - 1 \leq l_{1,a} < l_{2,a} < \cdots < l_{t-1,a} < l_{t+1,a} < \cdots. \]

Similarly, for each \( b \in \{0, 1, \cdots, M - 2\} \), let \( \{\tilde{l}_{b,k} \}_{k \geq 1} \) be a sequence of positive integers satisfying

\[ t \leq \tilde{l}_{1,b} < \tilde{l}_{2,b} < \cdots < \tilde{l}_{t,b} < \tilde{l}_{t+1,b} < \cdots. \]
The two families of sequences
\[ \{l_{k,a}\}_{k \geq 1, 0 \leq a \leq M-1} \text{ and } \{l_{k,b}\}_{k \geq 1, 0 \leq b \leq M-2} \]
are uniquely determined by the representation (5) of interval algorithm in Theorem 1. Details are found in [4]. The following is a corollary of Theorem 1.

**Corollary 1** For each \( x \in X \), we have
\[
p_x(x) = p_Y(x^{(t-1)}) \times \sum_{k \geq 1} \left[ \sum_{a=1}^{M-1} p_Y(a_{t+1}^{k+1} | a_{t-1}) + \sum_{b=0}^{M-2} p_Y(b_{t+1}^{k+1} | a_{t-1}) \right],
\]
where if \( l_{i,a} = t-1 \), then \( p_Y(a_{t}^{k+1+1} | a_{t-1}) = p_Y(a_t | a_{t-1}) \).

### 4. Performance Analysis of the Interval Algorithm

In this section we present a rigorous performance analysis of the interval algorithm using the expression of the interval algorithm we gave in the previous section. Set
\[
\eta_0(a, x | a_{t-1}) \triangleq \sum_{k \geq 1} p_Y(a_{t}^{k+1+1} | a_{t-1}), \quad \eta_1(b, x | a_{t-1}) \triangleq \sum_{k \geq 1} p_Y(b_{t}^{k+1+1} | a_{t-1}).
\]

Define two probability distributions on positive integers by
\[
p_Y^{(0)}(\cdot | a, x, a_{t-1}) \triangleq \frac{p_Y(a_{t}^{k+1+1} | a_{t-1})}{\eta_0(a, x | a_{t-1})} \quad \text{for } k = 1, 2, \ldots
\]
\[
p_Y^{(1)}(\cdot | b, x, a_{t-1}) \triangleq \frac{p_Y(b_{t}^{k+1+1} | a_{t-1})}{\eta_1(b, x | a_{t-1})} \quad \text{for } k = 1, 2, \ldots
\]

Let \( p_{\text{max}} \triangleq \max_{|a| < y}; p_{ij} \). Define the geometrical distribution \( p^* \) with parameter \( p_{\text{max}} \) by
\[
p^* \triangleq \left\{ p_{\text{max}}^{k-1}(1 - p_{\text{max}}) \right\}_{k = 1, 2, \ldots}
\]
For each \( a \in Y \), let \( Y(a) \) be a random variable having the distribution \( \{ \Pr(Y = y | Y = a) \}_{y \in Y} \). The entropy rate of \( \{ Y_i \}_{i=1}^{\infty} \) is given by
\[
H(Y_i | Y_i) = \sum_{y=0}^{M-1} p_Y(a) H(Y(a)).
\]
We set
\[
H_{\text{max}}(Y(a)) \triangleq \max_{a \in Y} H(Y(a)), \quad H_{\text{min}}(Y(a)) \triangleq \min_{a \in Y} H(Y(a)).
\]
The following is our main result.

**Theorem 2**
\[
\frac{H(X)}{H_{\text{max}}(Y(a))} \leq L \leq \frac{H(X)}{H_{\text{min}}(Y(a))} + \frac{\log 2(M - 1)}{\log 2 \Delta},
\]
where \( \Delta \) is a nonnegative number defined by
\[
\Delta \triangleq \sum_{x=0}^{N-1} \eta_0(a, x | a_{t-1}) \times \left\{ \sum_{a=1}^{M-1} \eta_1(b, x | a_{t-1}) \sum_{k=0}^{M-2} \eta_0(a, b | a_{t-1}) \right\}.
\]
By letting \( \Delta = 0 \) in (9), we obtain the upper bound of \( L \) derived by Han and Hoshi [3]. Hence our upper bound improves their one. The quantity \( \Delta \) indicates a lower bound of the deviation of the upper bound of \( L \) obtained by Han and Hoshi [3] from the true value of \( L \). This quantity is characterized with the \( p_Y \)-ary representation and the \( p_Y \)-ary co-representation of the endpoints of the intervals corresponding to the target random numbers.

### References


Random Number Generation Using Outputs from Multiple Beta Encoders

Koji Itaya† and Yuutaka Jitsumatsu‡

†Dept of Informatics
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
Email: itaya@me.kyushu-u.ac.jp, jitmatsu@inf.kyushu-u.ac.jp

‡Dept of Informatics
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
Email: itaya@me.kyushu-u.ac.jp, jitmatsu@inf.kyushu-u.ac.jp

Abstract—Beta encoder is an analog-to-digital converter which is robust to the fluctuation of the threshold voltage in a quantizer. Such a beta encoder is considered as a good candidate for random number generators (RNGs). In order to use a beta encoder as a RNG, strong correlation between consecutive bits must be eliminated. In this paper, the exclusive-or (EXOR) of outputs from multiple beta encoders is used as a random number. We investigated the statistical property of such a random number.

1. Introduction

Pseudo-random number generation is one of the most promising application of chaotic phenomenon observed in electronic circuits. There have been many researches on random number generator (RNG) using chaotic dynamics. Among them, RNGs using discrete-time chaotic dynamics with piecewise linear (PL) maps have attract much attention. Stojanovski and Kocarev [1] proposed to use PL map with two slopes $1 < k_1 < 2$ and $1 < k_2 < 2$. Using the same PL map as in [1] but with $k_1 = k_2 (= k)$, Addabbo et al. presented an interesting approach in which the amplification factor $k$ and a threshold are controlled by utilizing the observed statistics of the output binary codes [2]. Such a PL map with a slope $1 < k < 2$ is also used in beta encoders that is a kind of analog-to-digital (A/D) converters, where the slope is denoted by $\beta$ rather than $k$. In this paper, we consider a random number generation using exclusive-or (EXOR) of multiple $\beta$ encoders. The benefit of the proposed method over the existing ones [1][2] is that the proposed method is simple.

A $\beta$ encoder is an A/D converter, proposed by Daubechies et al. in 2002 [3]. A $\beta$ encoder aims to obtain $\beta$ expansion coefficients of input value $x$, where $\beta$ expansion of a real number $x \in (0, \frac{1}{\beta - 1}]$ is defined by $x = a_1/\beta + a_2/\beta^2 + a_3/\beta^3 + \ldots$, where $a_i \in [0, 1]$ are the expansion coefficients and $\beta$ is a fixed number in (1, 2). A $\beta$ expansion reduced to binary expansion if $\beta = 2$. The most important property of $\beta$ encoder is its robustness to the fluctuation of the threshold voltage value in the quantizer. Such a property enables us to use coarse precision capacitances, and low gain operational amplifiers [3]. Then, we can design a $\beta$ encoder extremely easily compared with other an A/D converter, and realize the miniaturization of a circuit area.

If we let a $\beta$ encoder output a large number of bits, such as ten thousand bits for one sample, then its bit sequence is considered a random number sequence. We observed an attractor in the hardware circuit [5]. We consider this attractor as a replacement for random physical phenomenon in the physical RNG. Thus we treat the output of a $\beta$ encoder as random number. However, a sequence of outputs of a $\beta$ encoder has a strong correlation between adjacent outputs. Therefore we need to make it closer to i.i.d. sequence by performing some post-processing.

In this paper, we consider a sequence of EXORs of plural outputs from $\beta$ encoders as random number sequence [6]. We performed a computer simulation, and evaluate the statistical properties of the generated random number sequences. A strong correlation between adjacent bits was highly suppressed by using multiple $\beta$ encoders.

2. Pulse Code Modulation and $\beta$ encoder

Pulse Code Modulation (PCM) is one of the standard analog-to-digital (A/D) conversion methods, which is based on binary expansion; the input analog value is converted into its binary expansion and then the binary expansion is expressed by a pulse train of its corresponding digital code. PCM consists of an amplifier that doubles the voltage accurately, a comparator that compares the voltage with the threshold value 1/2, and a subtractor circuit that reduces the voltage by a reference voltage 1. However, if the threshold in the comparator fluctuates a little from the value 1/2, then the output diverges. This observation suggests that we can not convert correctly.

A $\beta$ encoder is an A/D converter based on $\beta$-expansion, consisting of an amplifier with an amplification factor $\beta$ and a comparator with a threshold $\nu$. A circuit diagram of cyclic $\beta$ encoder is shown in Fig. 1. In a $\beta$ encoder, an ana-
log input value is converted into its corresponding \( \beta \)-ary expansion with finite precision. Let the output binary sequence obtained by a \( \beta \) encoder be \( \{a_i\}_{i=0}^{\infty} \), and initial input value be \( x = x_0 \in [0, \frac{1}{\beta-1}] \), then we have

\[
\begin{align*}
a_i &= Q_{\gamma}(\beta x_{i-1}), \quad i \geq 1, \\
x_i &= \beta x_{i-1} - a_i, \quad i \geq 1, \quad x_0 = x,
\end{align*}
\]

where \( Q_{\gamma}(x) \) is a comparator, defined by

\[
Q_{\gamma}(x) = \begin{cases} 
0 & 0 \leq x < \gamma/\beta, \\
1 & \gamma/\beta \leq x < 1/(\beta - 1).
\end{cases}
\]

The initial value \( x \) and \( \{a_i\}_{i=0}^{\infty} \) satisfy the following relation:

\[
x = \sum_{i=1}^{\infty} a_i \beta^{-i},
\]

where \( 1 < \beta < 2 \) and \( 1 \leq \gamma < \frac{1}{\beta-1} \). A \( \beta \)-expansion map is shown in Fig. 2. A \( \beta \)-expansion map is called greedy, cautious, and lazy maps, if \( \gamma = 1, \gamma = \frac{\beta}{2(\beta-1)} \), and \( \gamma = \frac{1}{\beta-1} \), respectively.

**Daubechies et al.’s flaky quantizer:** Daubechies et al. have proposed a model of quantizers having fluctuated threshold values, called a flaky quantizer [4]. A flaky quantizer is characterized by two threshold values, \( v_0 \) and \( v_1 \). If the voltage is less than \( v_0 \), the quantizer outputs 0, and if it is greater than \( v_1 \), the quantizer outputs 1, and if it is between \( v_0 \) and \( v_1 \), we do not know the quantizer outputs 0 or 1. The output of a \( \beta \) encoder with flaky quantizer is given by

\[
\begin{align*}
a_i &= Q^f_{\gamma,v_1}(\beta x_{i-1}), \quad i \geq 1, \\
x_i &= \beta x_{i-1} - a_i,
\end{align*}
\]

where \( Q^f_{\gamma,v_1}(x) \) is flaky quantizer defined by

\[
Q^f_{\gamma,v_1}(x) = \begin{cases} 
0 & x < v_0, \\
1 & x > v_1, \\
0 \text{ or } 1 & v_0 \leq x \leq v_1.
\end{cases}
\]

This \( Q^f_{\gamma,v_1}(x) \) is a model of a quantizer that outputs an incorrect judgment near the threshold. In the computer simulation in Section 4, we let \( Q^f_{\gamma,v_1}(x) \) to take 0 or 1 with equal probability if \( v_0 \leq \gamma \leq v_1 \). We define the map from \( x_i \) to \( x_{i+1} \) as \( C^f_{\beta,v_0,v_1}(x) \). Namely, we define (See Fig. 3.)

\[
C^f_{\beta,v_0,v_1}(x) = \beta x - Q^f_{\gamma,v_1}(\beta x).
\]

If the parameters \( v_0 \) and \( v_1 \) satisfy \( 1 \leq v_0 \leq v_1 \leq \frac{1}{\beta-1} \) and if the initial value satisfies \( 0 < x_0 < \frac{1}{\beta-1} \), then the orbit \( x_1, x_2, x_3, \ldots \) generated by \( x_i = C^f_{\beta,v_0,v_1}(x_{i-1}) \) does not diverge.

3. The proposed method

A \( \beta \) encoders can be realized in a very small CMOS circuit, therefore it is possible to implement multiple \( \beta \) encoders into one chip. Based on this fact, Hirata et al. have proposed to use EXOR of the outputs from multiple \( \beta \) encoders to generate a sequence of random binary numbers [6]. In this paper, we evaluate the performance of Hirata et al.’s method by using Daubechies’s flaky quantizers. We introduce the simulation method below.

**Fixed threshold model:** Firstly, we use quantizers with fixed thresholds to analyze Hirata et al.’s method. We assume that \( K \geq 1 \) \( \beta \) encoders with fixed thresholds are used. Once an analog signal is sampled, \( L \) bits of \( \beta \) expansion coefficients for this sample are obtained. Each of outputs of \( \beta \) encoders are distinguished by adding a superscript \((i)\), such as \( a_i^{(k)} \). An EXOR operation is performed among the outputs of \( \beta \) encoders of from 1 to \( K \), which is defined as \( b_i^{(K)} \), i.e.,

\[
b_i^{(K)} = a_i^{(1)} \oplus a_i^{(2)} \oplus \cdots \oplus a_i^{(K)}, \quad i = 1, 2, \ldots, L. \quad (7)
\]

Parameters for \( \beta \) encoders are set as follows: The \( \beta \) value is chosen from 1.7, 1.8 and 1.9. The threshold is chosen from three-values, greedy, cautious, and lazy. Initial value is randomly selected from \([0, \frac{1}{\beta-1}]\) with uniform distribution. The number of bits is \( L = 10,000 \).

**Flaky quantizer model:** Secondly, we use Daubechies’s flaky quantizers to analyze Hirata et al.’s method. Each of outputs of \( \beta \) encoders are denoted by \( \tilde{a}_i^{(k)} \). EXOR operation is performed among the outputs of \( \beta \) encoders from 1 to \( K \), which is defined as \( \tilde{b}_i^{(K)} \), i.e.,

\[
\tilde{b}_i^{(K)} = \tilde{a}_i^{(1)} \oplus \tilde{a}_i^{(2)} \oplus \cdots \oplus \tilde{a}_i^{(K)}, \quad i = 1, 2, \ldots, L. \quad (8)
\]

The initial value \( x_0^{(k)} \) is set to be \((v_0 + v_1)/2\beta\) for all \( k = 1, \ldots, K \). The reason why we choose this number as the
initial value is as follows: In a hardware $\beta$ encoder, the common mode voltage is one of the easiest voltages to be employed, which is modeled as $(v_0 + v_1)/2$. If the threshold of the quantizer in fixed, and if we start with the same initial value $x_0^{(k)}$, then the orbits $[x_0^{(k)}]$ should be the same for all of $k = 1, \ldots, K$. However, we employ the flaky quantizer here, so that the orbits $[x_0^{(k)}]$ for each $k$ are different. Therefore, we assume the same initial value for all $k$.

4. Experiment

We consider the two analytical models i.e., fixed threshold and flaky quantizer models, on the random number generation method using EXOR of outputs of plural $\beta$ encoders. The quality of the generated random number is evaluated by the following quantities: autocorrelation function, distribution of the sum of $b_1^{(K)}$, occurrence frequency of sequences of block length $N = 3, \ldots, 6$, and periodicity.

4.1. Autocorrelation function

Autocorrelation function of a random number sequence $b_1^{(K)}$, $i = 0, 1, \ldots, L - 1$, is defined by

$$R^{(K)}(\ell) = \frac{1}{L} \sum_{l=0}^{L-\ell-1} b_1^{(K)}(l) b_1^{(K)}(l+\ell).$$

(9)

Autocorrelation function of $b_1^{(K)}$ is determined similarly and denoted by $R^{(K)}(l)$. The number of $\beta$ encoders is set to be $K = 1, 2, 4$ in both models. Furthermore, we performed 300 independent trials in each simulations. The autocorrelation function for $\beta = 1.7$ is shown in Fig. 4. In this figure, "greedy" and "cautious" correspond to the fixed threshold model and others correspond to the flaky quantizer model. The "50\% flaky" means that the size of the range $[v_0, v_1]$ is 50\% of the full range $[1, 1/(\beta - 1)]$, while "flaky quantizer" means $[v_0, v_1] = [1, 1/(\beta - 1)]$. Autocorrelation functions for the lazy map are the almost the same as that of the greedy map and thus omitted here.

Fig. 4 shows that the autocorrelation values for $K = 1$ have strong correlations for the fixed threshold model. The autocorrelation function is close to the delta function if $K = 4$ for the fixed threshold model and if $K \geq 2$ for the flaky quantizer model.

Interestingly, we found that the autocorrelation function takes a positive value at $l = 1, -1$ for $\beta = 1.7$, but takes a negative value at $l = 1, -1$ for $\beta = 1.8$ and $1.9 K = 1$ for the fixed threshold model. If we employ the fixed threshold model, the autocorrelation function at $l = \pm 1$ must take a negative value for any threshold value. However, the hardware experiment by Tanaka et al. [7] reported the autocorrelation value takes a positive value at $l = \pm 1$. Namely, the fixed threshold model does not match the result of experiment by Tanaka et al. Using the flaky quantizer model, we can make $R^{(K)}(l)$ for $K = 1$ and $l = \pm 1$ to be negative. Hence, we consider the flaky quantizer model matches the hardware experiment more than the fixed threshold model.

4.2. Distribution of the sum of $b_1^{(K)}(i = 1, \ldots, L)$

The sum of $b_1^{(K)}$ from $i = 1$ to $L$ is evaluated for both fixed and flaky quantizer models. For an i.i.d. binary sequence $c_1, c_2, \ldots$, the central-limit theorem states that $Z_L = L^{-1/2} \sum_{l=1}^{L} c_l$ approaches a normal distribution as $L$ goes to infinity. We expect the distribution of the sum of $b_1^{(K)}$ is close to the normal distribution. We use the variational distance as an approximate measure. Let two distributions on a finite set $X$ be $P = (p_1, p_2, \ldots, p_K)$ and $Q = (q_1, q_2, \ldots, q_K)$. The variational distance between $P$ and $Q$ is defined by $d(P, Q) = \sum_{i=1}^{K} |p_i - q_i|$. The variational
We evaluate the number of occurrence of bit patterns such as (0, 0, ..., 0), (0, 0, ..., 1), ..., (1, 1, ..., 1) for block length of $N = 3, 4, 5$ and 6. The number of bits is $L = 120,000$. The variational distance between the empirical distribution of the bit patterns and uniform distribution for $N = 6$ and $\beta = 1.7$ is shown in Fig. 6. The variational was improved significantly at maximum value $K = 8$ in fixed threshold model, and $K = 2$ in flaky quantizer model. Fig. 6 shows that the quality of random numbers for cautious map is worse than the flaky quantizer. It is observed that the performance of 50% flaky quantizer is between those of the fixed threshold with cautious map and the flaky quantizer.

### 4.3. Occurrence frequency of bit patterns for short block length

We evaluate the number of occurrence of bit patterns such as (0, 0, ..., 0), (0, 0, ..., 1), ..., (1, 1, ..., 1) for block length of $N = 3, 4, 5$ and 6. The number of bits is $L = 120,000$. The variational distance between the empirical distribution of the bit patterns and uniform distribution for $N = 6$ and $\beta = 1.7$ is shown in Fig. 6. The variational was improved significantly at maximum value $K = 8$ in fixed threshold model, and $K = 2$ in flaky quantizer model. Fig. 6 shows that the quality of random numbers for cautious

### 4.4. Periodicity

We examined periodicity of the generated random numbers. We define $p \geq 1$ as the periodicity of an orbit $\{x_n\}_{n=0}^{\infty}$ if we find $x_n = x_{n+p}$ for some $n$. When we use multiple $\beta$ encoders, we define $p$ as the periodicity if we find $(x_1^{(1)}, x_2^{(1)}, ..., x_{L}^{(1)}) = (x_{1+n}^{(1)}, x_{2+n}^{(1)}, ..., x_{L+n}^{(1)})$ for some $n$. As far as we examined to length of sequence $L = 2^{20}$, we could not find such an example, and verified that the period of the generated bits is greater than $2^{20}$.

### 5. Conclusion

In this paper, we have evaluated performance of Hirata et al’s method by introducing Daubechies et al’s flaky quantizer. In the fixed threshold model, eight $\beta$ encoders are necessary to make the quality of generated sequence close to that of i.i.d. sequences. On the other hand, in the flaky quantizer model, two $\beta$ encoders were enough. These results show that the number of $\beta$-encoders required to attain a sufficient quality of the random number after taking EXOR of the plural of $\beta$ encoders is not so large as previously expected by Hirata et al.’s computer simulation employing the fixed threshold model.

### References


A Realization of Optimum Binary Spreading Sequences of Markov Chains Based on Discretized $\beta$-transformations

Hiroshi Fujisaki

Graduate School of Natural Science and Technology Kanazawa University Kakuma-machi, Kanazawa, Ishikawa, 920-1192 Japan
Email: fujisaki@ec.t.kanazawa-u.ac.jp

1. Introduction

Spreading sequences are a kind of pseudo-random numbers. It is one of the most crucial tasks in spread spectrum techniques to realize the optimum spreading sequences in terms of the performance of asynchronous spread-spectrum multiple-access (SSMA) communication systems where the sequences are used.

From the viewpoint of the performance of communication systems, it is bit error probabilities that are of the utmost importance as a measure of the reliability of the systems. The bit error probabilities in asynchronous SSMA communication systems were estimated by using Gaussian distributions whose variance was the average interference parameter (AIP) that was introduced by Pursley in [1] as a measure of the average signal-to-noise ratio (SNR) in asynchronous SSMA communication systems. This is the so-called standard Gaussian approximation (SGA).

Chaotic spreading sequences are the sequences of pseudo-random numbers generated by one-dimensional ergodic transformations, which is one of the applications of Ulam and von Neumann’s idea in [2]. It was found in [3] that a class of chaotic spreading sequences whose autocorrelations exponentially decay achieved a better performance in terms of the mean value of the AIP 1 as compared to Gold sequences whose auto-correlations are like a delta function. This discovery created a revolution in designing spreading sequences since sequences whose autocorrelations are like a delta function were commonly regarded as good sequences before the pioneering work. We note here that the chaotic sequences proposed in [3] are equivalent to the sequences generated by a class of Markov chains.

While Pursley defined the AIP as a measure of the average SNR in asynchronous SSMA communication systems, Yao pointed out in [4] that evaluations of bit error probabilities based on the SGA with the AIP were not valid for the systems with small numbers of users, low length of pseudonoise (PN) sequences, and high SNRs, which naturally posed the following questions: i) Why were evaluations of bit error probabilities based on the SGA with the AIP not valid for systems with small numbers of users and low lengths of PN sequences? ii) How can one give simple theoretical evaluations of bit error probabilities still valid for systems with small numbers of users and low lengths of PN sequences? These problems have often been discussed.

Motivated by the spreading sequences of Markov chains proposed in [3], we have studied to determine the optimum spreading sequences of Markov chains in terms of bit error probabilities in asynchronous SSMA communication systems. As a result of a series of studies [5]–[8], we have solved the above-mentioned Yao’s questions completely in virtue of the central limit theorem (CLT) together with large deviations analysis.

We showed that the SGA with the mean value of the AIP for estimations of bit error probabilities in such systems was the 0-th order approximation of the evaluation based on the CLT. As far as binary spreading sequences are concerned, correlational properties of the optimum spreading sequences in terms of the mean value of the AIP obtained in [9] coincide with the properties of the optimum sequences in terms of the bit error probabilities in the systems based on the CLT. We remark here that the result in [9] only gave correlational properties of the optimum spreading sequences. It did not tell us how to design the optimum spreading sequences in terms of the mean value of the AIP.

On the other hand, based on the CLT, we determined $k$-state Markov chains generating $k$-phase spreading sequences that minimize bit error probabilities in asynchronous SSMA communication systems in [6]. Moreover,
we found a novel class of spreading sequences, namely the phase-shift-free \( k \) (3)-phase spreading sequences, and showed in [7] that the optimum phase-shift-free \( k \)-phase spreading sequences of Markov chains were superior to the optimum binary spreading sequences of Markov chains in terms of the bit error probabilities in the system based on the CLT.

Unfortunately, however, such existence of the optimum sequences was theoretically determined and confirmed by using the piecewise-linear Markov transformations with the help of Monte-Carlo simulations. In fact, the optimum sequences are not available for practical use like Gold sequences because the idea in [2] requires handling real numbers in its applications. More precisely, the roundoff errors due to the truncation of real numbers occurs while iterating the Markov transformations by using computers.

Under these unpromising circumstances, a breakthrough was made in [10], where Bernoulli transformations were suggested for SSMA communication systems. Inspired by the results in [10], we defined discretized Markov transformations and found an algorithm to give the number of full-length sequences based on the discretized Markov transformations in [11].

In [12], we defined the piecewise-monotone-increasing Markov transformations, which included not only \( k \)-adic transformations but also Markov \( \beta \)-transformations. Besides, without knowing the total number of full-length sequences based on the discretized piecewise-monotone-increasing Markov transformations, we gave the bounded monotone truth-table algorithm for generating all full-length sequences which were based on the defined discretized Markov transformations.

In this report, we construct optimum binary spreading sequences of Markov chains in terms of bit error probabilities in asynchronous SSMA communication systems based on discretized \( \beta \)-transformations.

2. A Realization of Markov Chains with Prescribed Correlation Properties Based on \( \beta \)-transformations

In terms of bit error probabilities in asynchronous SSMA communication systems, the optimum \( k \) (2)-phase spreading sequences of Markov chains were determined in [7]. For the case where \( k = 2 \), the optimum binary spreading sequences of Markov chains are characterized by the sequence \( (Z_n)_{n=0}^{\infty} \) of \([1, -1]\)-valued stationary Markov chains with \( \mathbb{E}[Z_0] = 0 \) and \( \mathbb{E}[Z_0Z_1] = \left(-2 + \sqrt{3}\right)^\ell (\ell \geq 0) \). For a random variable \( Z \), we use \( \mathbb{E}[Z] \) to denote the expected value of \( Z \).

The correlation functions for sequences are measures of the similarity, or relatedness, between two sequences. Mathematically they are defined as follows.

**Definition 1** The normalized cross-correlation function of time delay \( \ell \) for the sequences \( X = (X_i)_{i=0}^{N-1} \) and \( Y = (Y_i)_{i=0}^{N-1} \) over \([1, -1]\) is defined by \( r_{xy}(\ell; X, Y) = 1/N \cdot \sum_{i=0}^{N-1} X_i Y_{i+\ell} \mod N, \) where \( \ell = 0, 1, \ldots, N-1 \) and, for integers \( a \) and \( b \) (\( \geq 1 \)), \( a \) (mod \( b \)) denotes the least residue of \( a \) to modulus \( b \). If \( X = Y \), we call \( r_{xy}(\ell; X, X) \) the normalized auto-correlation function, and simply denote it by \( r_{xx}(\ell; X) \).

In order to construct the sequence \( X \) with \( r_{xy}(\ell; X) = \left(-2 + \sqrt{3}\right)^\ell \), we recall the notion of Perron numbers defined in [13] as follows.

**Definition 2** The number \( \lambda \) is a Perron number if i) \( \lambda \) is a positive algebraic integer, and ii) \( \lambda > |\mu| \) for all other algebraic conjugates \( \mu \) of \( \lambda \). We use \( \mathbb{P} \) to denote the set of Perron numbers.

Let \( A \) be a non-negative integral matrix. If \( A^n > 0 \) for some positive integer \( n \), then \( A \) is called primitive, which is equivalent to irreducible and aperiodic. For an primitive matrix \( A \), we use \( A_n \) to denote the Perron-Frobenius eigenvalue of \( A \). Thus the Perron number is characterized by the following.

**Theorem 1** (Lind [13]) \( \lambda \in \mathbb{P} \) iff \( \lambda = \lambda_n \) for some primitive \( A_n \).

For our purpose, since the correlation function in question has only one parameter, it suffices to consider \( A \in \mathbb{P} \) with degree 2. The minimal polynomial of \( \lambda \) over \( \mathbb{Q} \) is defined by \( f(t) = t^2 - c_1 t - c_2 \) where \( c_1, c_2 \in \mathbb{Z} \). Its companion matrix of is given by \( B = \begin{pmatrix} 0 & c_2 \\ 1 & c_1 \end{pmatrix} \). Recall that the characteristic polynomial and the minimal polynomial of \( B \) are equal to \( f(t) \). In order to associate a \( \beta \)-transformation with \( B \), in what follows, we assume \( 0 < c_2 \leq c_1 \).

The adjacency matrix \( A \) of the \( \beta \)-transformation \( T \) associated with the above companion matrix \( B \) is given by

\[
A = \begin{pmatrix}
1 & \cdots & 1 & 1 & \cdots & 1 \\
\vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\
1 & \cdots & 1 & 1 & \cdots & 1 \\
1 & \cdots & 1 & 1 & \cdots & 1 \\
\cdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
1 & \cdots & 1 & 1 & \cdots & 1 \\
\end{pmatrix}
\begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.
\]

For almost every \( x \) in \([0, 1]\), the \( n \)-th iterate \( T^n(x) \), where \( T^0(x) = x \) and \( T^n(x) = T^{n-1}(T(x)) \) for \( n = 1, 2, \ldots \), together with a map \( \Psi : [0,1] \to [1,-1] \) defined by \( \Psi(x) = 1 \) if \( x < c_1/\beta \) and \( \Psi(x) = -1 \) otherwise, generates a sequence \((Z_n)_{n=0}^{\infty}\) of \([1, -1]\)-valued Markov chain by setting \( Z_n = \Psi(T^n(x)) \). Thus we obtain \( \mathbb{E}[Z_0Z_1] = \left(\frac{1 + \bar{\lambda}}{1 - \lambda} \right)^\ell - 4\bar{\lambda}(\lambda - \bar{\lambda})^2 \left(\frac{1}{\lambda} \right)^\ell (\ell \geq 0) \), where \( \bar{\lambda} \) is the algebraic conjugate of \( \lambda \), the unique integral solution \((c_1, c_2)\) of an equation \(-2 + \sqrt{3} = \lambda/\bar{\lambda} = c_1 - \sqrt{c_1^2 + 4c_2} \) with \( 0 < c_2 \leq c_1 \).
is given by $c_1 = c_2 = 2$. Eventually, we obtain the $\beta$-transformation $T$ with $\beta = 1 + \sqrt{3}$ which is the positive solution of $t^2 - 2t - 2 = 0$. The graph of $T$ is given in Fig. 1.

![Figure 1: The $\beta$-transformation with $\beta = 1 + \sqrt{3}$](image)

Although we successfully obtain a sequence of $\{1, -1\}$-valued stationary Markov chain with $\mathbb{E}[Z_n Z_0] = \left(-2 + \sqrt{3}\right)^t (t \geq 0)$, we still have $\mathbb{E}[Z_n] = (\lambda + 1)/(\lambda - 1) = 1/\sqrt{3} \neq 0$ since the stationary distribution of the chain is given by $(p_1, p_2) = 1/(\lambda - 1) : (-\lambda, \lambda) = 1/(2 \sqrt{3}) \cdot (-1 + \sqrt{3}, 1 + \sqrt{3})$, which is not uniform.

In the next section, without changing the realized correlated properties of the binary optimum spreading sequences of Markov chains, we transform the distribution $(p_1, p_2)$ of the sequences by virtue of sliding block codes.

3. A Realization of Markov Chains with Prescribed Correlation Properties with the Uniform Distribution Based on Discretized $\beta$-transformations

Let $\Sigma$ be a finite alphabet. The full $\Sigma$-shift is denoted by $\Sigma^\mathbb{N} = \{x = (x_i)_{i \in \mathbb{Z}} : x_i \in \Sigma\}$ which is endowed with the product topology arising from the discrete topology on $\Sigma$. The shift transformation $\sigma : \Sigma^\mathbb{N} \to \Sigma^\mathbb{N}$ is defined by $\sigma(x_i)_{i \in \mathbb{Z}} = (x_{i+1})_{i \in \mathbb{Z}}$. The closed shift-invariant subsets of $\Sigma^\mathbb{N}$ are called subshifts.

We call elements $u = u_1 u_2 \cdots u_n \in \Sigma^n$ blocks over $\Sigma$ of length $n \ (n \geq 1)$. We use $e$ to denote the empty block. For a subshift $X$, we use $L_n(X)$ to denote the collection of all $n$-blocks appearing in points in $X$. The language of $X$ is the collection $L(X) = \bigcup_{n=0}^{\infty} L_n(X)$, where $L_0(X) = \{e\}$.

A shift of finite type (SFT) is a subshift that can be described by a finite set of forbidden blocks. For a given finite set $F$ of forbidden blocks, we use $X_F$ to denote the SFT.

The symbolic representation of $\beta$-expansions of real numbers with $\beta = 1 + \sqrt{3}$, which is realized by the itterates of the $\beta$-transformation $T$ shown in Fig. 1, is given by the SFT $X_F \subset \Sigma^\mathbb{N}$ where $\Sigma = \{0, 1, 2\}$ and $F = \{2\}$. Its graph representation $G$ is given in Fig. 2 which also represents $T$.

Setting $G = G^{(2)}$, we obtain a sequence $(G^{(n)})_{n=2}^\infty$ of higher edge graphs of $G$. For each $n \geq 2$, we use $H_n$ to denote the Eulerian subgraph spanning $G^{(n)}$ with maximal number of edges, whose Eulerian circuits are the full-length sequences based on the discretized $\beta$-transformation $T$ with $\beta = 1 + \sqrt{3}$. In Fig. 2, we see that $G$ is Eulerian. Thus we have $G = G^{(2)} = H_2$ in this case. In the Eulerian subgraph $H_2$, we obtain a full-length sequence $001021120$ for instance. The length $|B_n|$ of full-length length sequences is given by $|B_n| = \beta^n + 1/\beta (n \geq 2)$ in [14], where $\beta = 1 - \sqrt{3}$, which is the algebraic conjugate of $\beta$.

Now we are in the position to construct the optimum binary spreading sequences of Markov chains based on the discretized $\beta$-transformations with $\beta = 1 + \sqrt{3}$.

A total order relation $\leq$ on $L(X_F) \setminus \{e\}$ is defined by the following: for any $u = u_1 u_2 \cdots u_m \ (m \geq 1)$ and $v = v_1 v_2 \cdots v_n \ (n \geq 1)$ in $L(X_F)$, $u \leq v$ if and only if

$$\frac{u_1}{\beta} + \frac{u_2}{\beta^2} + \cdots + \frac{u_m}{\beta^m} \leq \frac{v_1}{\beta} + \frac{v_2}{\beta^2} + \cdots + \frac{v_n}{\beta^n}.$$

For simplicity, we use $L$ to denote the length $|B_n|$ of full-length length sequences. We define a block code $\Phi : \{0, 1, 2\}^L \to \{1, -1\}$ by for $v = v_1 v_2 \cdots v_L \in \{0, 1, 2\}^L$,

$$\Phi(v) =\begin{cases} 1 & \text{if } v \leq 0, \\ -1 & \text{if } 0 < v \leq 2, \\ 1 & \text{if } v > 2. \end{cases}$$

We use $S$ to denote the shift transformation on $\{0, 1, 2\}^L$, i.e., $S(v_1, v_2, \cdots, v_{L-1}, v_L) = (v_2, v_3, \cdots, v_L, v_1)$ for $v = v_1 v_2 \cdots v_L \in \{0, 1, 2\}^L$. Thus we obtain a sliding block code $\phi$ for periodic sequences of period $L$ defined by $\phi(v^n) = (\Phi(v) \Phi(S(v)) \Phi(S^2(v)) \cdots \Phi(S^{L-1}(v)))^\infty$, where $u^n = \cdots uu \cdots$ for a block $u$. The sliding block code $\phi$ transform the full-length sequence over $\Sigma = \{0, 1, 2\}$ based on the discretized Markov $\beta$-transformation with $\beta = 1 + \sqrt{3}$ into the optimum binary spreading sequences of Markov chains as follows.

Let $X$ be a full-length sequence over $\Sigma = \{0, 1, 2\}$ of length $L = |B_n|$ based on the discretized Markov $\beta$-transformation with $\beta = 1 + \sqrt{3}$. Thus the optimum binary spreading sequence of Markov chain is realized by $Y = \Phi(X) \Phi(S(X)) \Phi(S^2(X)) \cdots \Phi(S^{L-1}(X))$. We here give an example of the optimum binary spreading sequences of Markov chains of length $|B_n|$.

**Example 1** For $n = 3$, we have $L = 20$ and

$$001002011012102112 \xrightarrow{\phi^{(2)}} 11101011001010010001,$$
where in the right hand side, we use 0 to denote −1 for simplicity.

Applying the previous results in [15] to the optimum binary spreading sequences of Markov chains, we have

**Theorem 2** For 0 ≤ ℓ ≤ n − 1, we obtain

\[
r_{\beta^n}(\ell; Y) = (-2 + \sqrt{3})^\ell + \left\{ \frac{\beta}{\beta'} \right\}^\ell - \left\{ \frac{\beta}{\beta'} \right\}^n \left\{ 1 + \left( \frac{\beta}{\beta'} \right)^\ell \right\}.
\]

This implies \( r_{\beta^n}(\ell; Y) = \mathbb{E}[Z_{\ell}Z_1] + O\left( \frac{\beta}{\beta'} \right)^n \), where O is the big O notation from the Landau symbol.

### 4. Experimental Results

A short table of values of the length \( |\mathcal{B}_n| \), the total number \( v_n \) of the full-length sequences over \{0, 1, 2\} in \( H_n \), and the total number \( v_{\beta} \) of the realized optimum binary spreading sequences of Markov chains are given in Table 1, respectively.

<table>
<thead>
<tr>
<th>( n )</th>
<th># of seq.s</th>
<th># of seq.s w/ uniform dist.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>1728</td>
</tr>
</tbody>
</table>

Fig. 3 shows the theoretical estimations based on the CLT given in [5] and the experimental results of bit error probabilities in asynchronous SSMA communication systems using the realized optimum binary spreading sequences of Markov chains as a function of the number of users \( J \) for \( N = 56 \). In this figure, the experimental results and the theoretical estimations based on the CLT agree properly with each other.

**Figure 3:** The bit error probabilities.

### Acknowledgments

This study was supported by the Telecommunications Advancement Foundation. The author is grateful to Mr. Jinse Shimo for help with the computer simulations.

### References


Synchronization in a Coupled Izhikevich Neuron Model

Kaede Naiki, Yutaka Shimada, Kantaro Fujiwara and Tohru Ikeguchi

†Department of Management Science, Graduate School of Engineering, Tokyo University of Science
‡Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
Niijuku, Katsushika-ku, Tokyo, Japan
Email: naiki@hisenkei.net

Abstract—We investigated a response in a gap-junctionally coupled Izhikevich neuron model. We used a diversity index of inter spike intervals (ISIs) to evaluate the response. As a result, we discovered that periodically firing neurons exhibit chaotic firing by the coupling. In addition, we investigated a synchronization phenomenon in the coupled neurons.

1. Introduction

Izhikevich neuron model[1] is represented as follows:

\[
\begin{align*}
\dot{v} &= 0.04v^2 + 5v + 140 - u + I, \\
\dot{u} &= a(bv - u).
\end{align*}
\]  

(1)

When \( v \geq 30 \), \( v \leftarrow c \), \( u \leftarrow u + d \). In the model, \( v \) represents membrane potential, \( u \) represents recovery potential, and \( a \), \( b \), \( c \), \( d \) and \( I \) represent parameters. The parameter \( I \) describes total synaptic currents into a neuron. This model can reproduce various types of neuronal firings.

In our brain, more than ten billion neurons mutually connect. There are two types of neuronal couplings: chemical couplings and electrical couplings. The electrical couplings are called gap-junction. It is well known that the gap-junction plays an important role for realizing synchronization in our brain [2][3], however, it is unclear how chaotic firing neuron changes its response by the gap-junction. In this paper, we coupled two Izhikevich neurons with the gap-junction. Then, the coupled neurons by the gap-junction can be represented as follows:

\[
\begin{align*}
\dot{v}_i &= 0.04v_i^2 + 5v_i + 140 - u_i + I_i + \sum_{j \neq i} w_{ij}(v_j - v_i), \\
\dot{u}_i &= a_i(b_iv_i - u_i).
\end{align*}
\]  

(2)

In Eq. (2), \( w_{ij} \) represents coupling coefficients. In this paper, the number of coupled neurons is two, then \( N = 2 \). In this paper, the coupling coefficients \( w_{ij} \) and \( w_{ji} \) are the same value, namely \( w_{ij} = w_{ji} = w \) and we coupled two neurons which have the same parameter values, namely \( a_1 = a \), \( b_1 = b \), \( c_1 = c \), \( d_1 = d \) and \( I_1 = I \) for all \( i \).

2. Chaotic response in a single neuron

2.1. Methods

The Izhikevich neuron model can reproduce chaotic response with the parameter values of \( a = 0.2 \), \( b = 2.0 \), \( c = -55 \), \( d = 2 \) and \( I = -99 \)[5]. In these parameter values, the parameter \( I = -99 \), which means that synaptic currents is negative. Then, we investigated whether we could find chaotic response with positive \( I \). To investigate parameter regions of chaotic response, we used the Lyapunov exponent, which is a measure of chaos. When the Lyapunov exponent \( \lambda \) is positive, the response is chaotic. To estimate the Lyapunov exponent from the discontinuous systems, we used a return map from the Poincaré map. We set the Poincaré section \( \Sigma \) as \( \Sigma = v = c \), namely, \( v = 30 \). Therefore, the Poincaré map can be defined as follows:

\[
u_{i+1} = T_{\Sigma}(u_i),
\]  

(3)

where \( u_i \) represents the \( i \)-th value of the map. Then, the Lyapunov exponent \( \lambda \) can be defined as follows:

\[
\lambda = \frac{1}{n} \sum_{i=1}^{n} \log |g'(u_i)|.
\]  

(4)

In Eq. (4), \( n \) indicates the total number of firings, \( g(u_i) \) indicates the function of the return map. Figure 1 shows the return map form of the Poincaré map, namely, \( g(u_i) \) with the parameter values of \( a = 0.0255 \), \( b = 0.3 \), \( c = -50 \), \( d = 2 \) and \( I = 10 \).

2.2. Results

In Fig. 2, we estimated the Lyapunov exponents in case that the value of the parameter \( a \) is varied from 0.01 to 0.08. Other parameter values are fixed as follows: \( b = 0.3 \), \( c = -50 \), \( d = 2 \) and \( I = 10 \).

As shown in Fig. 2, Lyapunov exponent \( \lambda \) takes a positive value around \( a = 0.03 \). Therefore, we could find a parameter range that the neuron exhibits chaotic response even though the parameter \( I \) takes a positive value.

Figure 3 shows a time series of the \( v \) with the parameter values of \( a = 0.0317 \), \( b = 0.2 \), \( c = -50 \), \( d = 2 \) and \( I = 10 \). From Fig. 3, we can see that the state of neuron exhibits periodic responses with five firings. Figure 4 shows the trajectory in the \((u, v)\) plane and the time series of the \( v \) with
Figure 1: Return map form of the Poincaré map. The horizontal axis is $u_i$ of the map and the vertical axis is $u_{i+1}$ ($a=0.0255$, $b=0.3$, $c=-50$, $d=2$ and $I=10$).

Figure 2: Result of estimating the Lyapunov exponents. The horizontal axis is the parameter $a$ and the vertical axis is the Lyapunov exponent ($b=0.3$, $c=-50$, $d=2$ and $I=10$).

Figure 3: Time series of the membrane potential $v$ of a single neuron in a periodic parameter values ($a=0.0317$, $b=0.3$, $c=-50$, $d=2$ and $I=10$).

Figure 4: Response of a single neuron with a chaotic parameter values ($a=0.032$, $b=0.3$, $c=-50$, $d=2$ and $I=10$) of (a) trajectory in the $(u,v)$ plane. The horizontal axis is membrane potential $v$ and the vertical axis is recovery potential $u$, and (b) a time series of the membrane potential $v$.

3. Synchronization in a coupled neuron

In this section, we investigated how neuronal responses change by coupling two neurons with the gap-junction as shown in Eq. (2). We connected two neurons with the parameter values of $a=0.03$, $b=0.3$, $c=-50$, $d=2$ and $I=10$. With these parameters, the Lyapunov exponent $\lambda>0$. In this case, the neuron model exhibits chaotic response as a single neuron (Fig. 5(a)). Fig. 5(b) shows the responses of coupled neurons with the coupling coefficient $\omega=-0.05$. Although the response of a single neuron is chaotic (Fig. 5(a)), when two neurons are coupled, the responses of two neurons become periodic with the period five. Moreover, responses of two neurons exhibit anti-phase synchronization.

Figure 6 shows the response of coupled neurons with the parameters $a=0.0647$, $b=0.2$, $c=-55$, $d=2$ and $I=10$. Figure 6(a) shows the trajectory of $(u,v)$ of a single neuron. Figure 6(b), 6(c) and 6(d) shows the responses of two coupled neurons with the coupling coefficient $\omega=0.05$. As shown in Fig. 6(a), the response of a single neuron is period two, when $a=0.0647$, $b=0.2$, $c=-55$, $d=2$ and $I=10$. On the other hand, as shown in Fig. 6(b) and 6(c), when the coupling coefficient is $\omega=0.05$, the trajectories of $(u,v)$ are chaotic. Figure 6(d) shows the time series of membrane potential $v$ of two neurons. From Fig.
6(d), we can observe that membrane potential of two neurons have a fixed phase difference, although their responses are chaotic. Namely, in this case, coupled two neurons changed their responses from periodic to chaotic by the gap-junction. Moreover, the coupled two neurons exhibit chaotic synchronization.

![Time series of the membrane potential](image)

**Figure 5:** Time series of the membrane potential \( v \) of (a) a single neuron without coupling and (b) coupled two neurons with the coupling coefficient \( w = -0.05 \). Parameter values are set to \( a = 0.03, b = 0.3, c = -50, d = 2 \) and \( I = 10 \), with these parameters, the neuron exhibits chaotic response.

4. Evaluation by diversity index of ISI

4.1. Methods

To evaluate connected neuronal responses of coupled neurons, we used a diversity index of ISI\(^6\), which is defined as follows:

\[
R = \frac{n}{m},
\]

where \( n \) indicates the number of types of ISIs and \( m \) indicates the total number of ISIs. When the difference between the \( i \)-th ISI \( (i = 1, 2, \ldots, m) \) \( t_i \) and the \( j \)-th ISI \( (j = 1, 2, \ldots, m) \) \( t_j \) is larger than \( 1.0 \times 10^{-15} \), that is, \( |t_i - t_j| > 1.0 \times 10^{-15} \), \( t_i \) and \( t_j \) are defined to be different. If neuron exhibits irregular, possibly chaotic response, \( n \) becomes large then \( R \rightarrow 1 \). In contrast, if neuron exhibits periodic response, \( R \rightarrow 0 \). In this paper, we measured the diversity index of ISI for one neuron even though two neurons are coupled.

4.2. Results

We evaluated the diversity index of ISI when parameter values are \( b = 0.2, d = 2 \) and \( I = 10 \). (Fig. 7). In Fig. 7, we changed two parameters \( a \) and \( c \) in the range of \( 0.01 \leq a \leq 0.07 \) and \( -65 \leq c \leq -45 \). The coupling coefficient is set to \( w = 0.1 \) in Fig. 7(a) and \( w = 0.05 \) in Fig. 7(b). From Fig. 7, bifurcation structures of two cases look similar. However, in the range of \( a = 0.04, -60 < c < -56 \), the values of diversity indexes of ISI show different values. From these results, the responses of two neurons are changed by the coupling coefficient.

Next, let us focus on the parameter values of \( a = 0.04, b = 0.2, c = -58.8, d = 2 \) and \( I = 10 \). In these parameter values, the diversity index of ISI is \( R = 0.748 \) when the coupling coefficient \( w = 0.1 \), as shown in Fig. 7(a), that is, the coupled neurons seem to show chaotic response. On the other hand, the diversity index of ISI is \( R = 0.0096 \) when the coupling coefficient \( w = 0.05 \) as show in Fig. 7(b), that is, the coupled neurons shown periodic response. Figure 8(a) shows the time series of \( v \) of a single neuron when \( a = 0.04, b = 0.2, c = -58.8, d = 2 \) and \( I = 10 \). Figure 8(b) and 8(c) show the time series of \( v \) of two coupled neurons when \( w = 0.05 \) (Fig. 8(b)) and \( w = 0.1 \) (Fig. 8(c)). As shown in Fig. 8(a), the response of a single neuron is periodic. Then, as shown in Fig. 8(b), the responses of coupled neurons are also periodic when \( w = 0.05 \). In Fig. 8(b) and 8(c), we coupled the neurons with the same parameter values as Fig. 8(a). However, as shown in Fig. 8(b), two neurons after the coupling exhibit different periodic responses compared with Fig. 8(a). On the other hand, as shown in Fig. 8(c), coupled neurons exhibit chaotic responses. As a result, it is indicated that periodically firing neuron can exhibit chaotic firing by changing the coupling strength.
5. Conclusion

In this paper, we investigated the chaotic response and synchronization observed in the Izhikevich neuron[1]. We discovered that positive synaptic currents ($I > 0$) that lead to chaotic response exist. Moreover, we coupled two Izhikevich neurons. By changing coupling coefficients, neurons can change their responses and be synchronized. To clarify the mechanism of synchronization of coupled neurons, we should investigate other synchronization phenomena in detail.

Acknowledgment

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. was supported by Grant-in-Aid for Research Activity Startup (No. 26880020) from JSPS. The research of K.F. was supported by Grant-in-Aid for Challenging Exploratory Research (No. 15K12137) and by Grant-in-Aid for young Scientists(B) (No. 16K16138) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Challenging Exploratory Research (No. 24650116) from JSPS and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No.15KT0112) from JSPS.

References

Analysis of Synaptic Dynamics during Infra-Slow Oscillation

Toshihiro Kobayashi\(^{\dagger}\), Yutaka Shimada\(^{\dagger, \ddagger}\), Kantaro Fujiwara\(^{\dagger, \ddagger}\) and Tohru Ikeguchi\(^{\dagger, \ddagger}\)

\(^{\dagger}\)Department of Management Science, Graduate School of Engineering, Tokyo University of Science
\(^{\ddagger}\)Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6–3–1 Niijuku, Katsushika-ku, Tokyo, Japan

Abstract—In our brains, interaction between neurons generates a variety of rhythms, for example, alpha, delta and gamma rhythms, and so on. These rhythms are observed from mathematical neural network models with spike-timing-dependent plasticity (STDP). We have already discovered that a mathematical model with the STDP can reproduce a rhythm of very low frequency, or infra-slow oscillation. In this paper, we analyzed the synaptic dynamics during the infra-slow oscillation. As a result, it is indicated that synaptic dynamics plays a key role for reproducing the infra-slow oscillation.

1. Introduction

Billions of neurons exist in our brains and their interaction generates a variety of rhythms. Among them, one of the most interesting rhythms is infra-slow oscillation (ISO). ISO was discovered by Aladjalova\(^{[1]}\) with a local field potential recorded from the rabbit neocortex. Although ISO has been observed in various kinds of mammalian brains\(^{[2]}\), its generation mechanism remains unknown. On the other hand, delta rhythms (2–4 [Hz]) and gamma rhythms (30–100 [Hz]) are reproduced by a mathematical neural network model with axonal conduction delays and spike-timing-dependent plasticity (STDP)\(^{[3]}\).

In our former study, we investigated a neural mechanism to reproduce ISO, conducting numerical simulations with changing the curvature of the STDP function\(^{[4]}\). In this study, we investigated the spike timing difference of presynaptic and postsynaptic neurons and temporal change in the synaptic weights of all synapses.

2. STDP learning

We used the STDP rule for learning of the neural network. In the STDP, the magnitude of change rates in synaptic weights depends on the timing of spikes: if a presynaptic spike arrives at the postsynaptic neuron before the postsynaptic neuron fires, the synapse is potentiated (long-term potentiation, LTP). If the presynaptic spike arrives at the postsynaptic neuron after the postsynaptic neuron fired, the synapse is depressed (long-term depression, LTD). The magnitude of the change in synaptic weights is decided by STDP function which is represented as follows\(^{[5]}\):

\[
\Delta w_{ij}(\Delta t_{ij}) = \begin{cases} A_+ \exp(-\frac{\Delta t_{ij}}{\tau_+}) (\Delta t_{ij} > 0), \\ -A_- \exp(\frac{\Delta t_{ij}}{\tau_-}) (\Delta t_{ij} < 0), \end{cases}
\]

where \(\Delta t_{ij} = t_i - t_j - \delta_{ij}\), \(t_i\) is the firing time of postsynaptic neuron \(i\), \(t_j\) is the firing time of presynaptic neuron \(j\), \(\delta_{ij}\) is conduction delay from neuron \(j\) to neuron \(i\), \(A_+\) is the maximum value of LTD, \(A_-\) is the maximum value of LTD, and \(\tau\) is the time constant of LTP and LTD.

3. Methods

The neural network we used consists of 1,000 randomly connected Izhikevich neuron\(^{[3]}\). We prepared 800 excitatory neurons and 200 inhibitory neurons. In this paper, we used regular spiking neurons for the excitatory neurons, and fast spiking neurons for the inhibitory neurons. Each neuron has 100 synapses connected to other neurons. Every excitatory neuron is connected to 100 neurons that are randomly chosen from all neurons, while every inhibitory neuron is connected to 100 neurons that are randomly chosen from excitatory neurons.

Conduction delays among all neurons are random integers between 1 [ms] and 20 [ms]. The excitatory connection obeys the STDP learning rule with every 1 second. The maximum value of LTP, \(A_+\), is 0.1 and the maximum value of LTD, \(A_-\), is 0.12. The initial values of the weights are set to 6, the maximum value is limited to 10, and the minimum value is limited to 0. The change of the synaptic weights adopts the nearest-neighbor spiking. The excitatory connections are updated every second by Eq. (2):

\[
w_{ij}(t) = w_{ij}(t-1) + \sum_{t_i=t-1}^{t} \Delta w_{ij}(\Delta t_{ij}), \tag{2}
\]

where \(\Delta w_{ij}(\Delta t_{ij})\) is defined by Eq. (1) and depends on \(t_i\) and \(t_j\). Inhibitory connection weights are fixed to \(-5\). A randomly chosen neuron receives a pulse of 20 [mV] every 1 [ms] as a random thalamic input. With these experimental conditions, we changed the value of the parameter \(\tau\) which determines the curvature of
the STDP function in Eq. (1), and investigated the time series of firing rates and synaptic weights.

4. Results

4.1. Temporal Change of Firing Rates and Synaptic Weights

Figure 1: Temporal changes of firing rates and average synaptic weights\[4\]. The horizontal axis is time [s], the left vertical axis is the synaptic weight and the right vertical axis is the firing rate [Hz]. When $\tau = 1$, the firing rate and the synaptic weight oscillate with very slow rhythms. Both excitatory to excitatory and excitatory to inhibitory synaptic weights take almost the same values.

Figure 1 shows a temporal change of firing rates and average synaptic weights when $\tau = 10$ and 1. We defined the firing rate as the average firing frequency of a single neuron among all neurons every one second. Namely, we defined the firing rate by $m/N$ [Hz], when $m$ firings are observed from $N$ neurons per second. The average synaptic weight is the average value of synaptic weights of all connections including excitatory and inhibitory connections in every second.

As shown in Fig. 1(a), when $\tau = 10$, the firing rate fluctuated with high frequency with almost constant amplitude. As shown in Fig. 1(b), when $\tau = 1$, the firing rate repeated sudden rise and fall with very slow frequency. This tendency was observed when $\tau \sim 1$.

Focusing on synaptic weights, when $\tau = 10$, the synaptic weights are constant as shown in Fig. 1(a). The synaptic weights between excitatory and inhibitory neurons are stronger than that of excitatory and excitatory neurons.

On the other hand, as shown in Fig. 1(b), the synaptic weights oscillate with the same period as the firing rate when $\tau = 1$. The synaptic weights from excitatory to inhibitory neurons became smaller than that from excitatory to excitatory neurons. The difference between excitatory–inhibitory synaptic weights and excitatory–excitatory synaptic weights is smaller when $\tau = 1$ than when $\tau = 10$.

4.2. Temporal Change of Synaptic Weights

Figure 2: Temporal change of all synaptic weights selected from excitatory neuron arbitrarily. As shown in Fig. 2, when $\tau = 10$, the synaptic weights are almost constant and are separated into maximum and minimum values. However, when $\tau = 1$, the synaptic weights change with almost the same period as the firing rate. Some synapses have intermediate values as well as maximum and minimum values.

4.3. Histogram of Spike Timing Difference

To analyze the relation between the firing of presynaptic neuron and the firing of postsynaptic neuron, we investigated the spike timing differences between presynaptic and postsynaptic neuron. In this paper, we defined spike timing difference as $\Delta t_{ij}$ in Eq. (1). Figure 3(a) shows the frequency distribution of the spike timing difference when $\tau = 10$. As shown in the
figure, the frequency of the spike timing difference repeats high and low values periodically. This rhythm is about 70 Hz (gamma rhythm). The appearance of the gamma rhythm has already been shown in the experiment by Izhikevich [3]. This tendency does not depend on time. Therefore, result of Fig. 3(a) is consistent with the experiment by Izhikevich [3].

**Figure 3:** Frequency distributions of spike timing differences between presynaptic and postsynaptic neurons when (a) $\tau = 10$, (b) $\tau = 1$ during low firing rate and (c) $\tau = 1$ during high firing rate. The horizontal axis is difference of spike timing. The vertical axis is frequency.

Figure 3(b) and 3(c) show the histograms of the spike timing differences when $\tau = 1$. Figure 3(b) is the frequency distribution when the firing rate is low, and Fig. 3(c) is the frequency distribution when the firing rate is high. As shown in the figures, there exist more spike timing differences of negative values than those of positive values. It means that LTD occurs more frequently than LTP. In particular, when the high firing rates appear, the number of LTD increase extremely (Fig. 3(c)).

**4.4. Change of Synaptic Weights Which Induce Neuronal Firing**

**Figure 4:** Temporal change of synaptic weights which induced neuronal firing. The horizontal axis is time [s]. The vertical axis is the synaptic weights. (a) shows the change when $\tau = 10$ and (b) shows the change when $\tau = 1$. When $\tau = 1$, the synaptic weights become stronger as the firing rates.

When $\tau = 1$, there are a large number of LTD during high firing rate. We investigated the reason why the firing rate becomes high despite the large number of LTD.

Each neuron receives a number of inputs through synapses. However, such inputs do not always make the neuron fire. We focused on the synapses which induce firings of neurons. We plotted the time series of the average synaptic weight in such synapses. Figure 4 shows the temporal change of the average value of the synaptic weights in the synapses which induced neuronal firing. As shown in the figure, when $\tau = 10$, the synaptic weight fluctuates almost stationarily. When $\tau = 1$, the synaptic weight becomes stronger near 10500 [s], and this corresponds to sudden rise of the firing rates as shown in Fig. 1(b). From this result, it is revealed that firing rates can be high by inputs from specific synapses even if there are more LTD than LTP in average.

**4.5. Histogram of Synaptic Weights**

**Figure 5** shows the histogram of synaptic weights. As shown in the figure, when $\tau = 10$, the synaptic weights are almost completely separated into maximum and minimum values. On the other hand, when $\tau = 1$, the synaptic weights are not completely divided...
into maximum and minimum value and there are more
synapses which have intermediate values.

Figure 5: Histograms of synaptic weights. The horizontal axis is synaptic weights. The vertical axis is frequency. The synaptic weights are almost completely separated into maximum and minimum values when \( \tau = 10 \), while the synaptic weights are not completely separated into maximum and minimum values when \( \tau = 1 \).

5. Mechanism to Generate ISO

The oscillation of the firing rate shown in Fig. 1(b) is explained by the dynamics of synaptic weights. When the value of \( \tau \) is large, the STDP learning window has a large width along the temporal direction. The STDP learning with such a wide window has many chances for learning and the synaptic weights are separated into maximum and minimum values. Because the bimodal distribution does not change its shape easily by LTP and LTD, the learning converges and the firing rate becomes stable.

On the other hand, when the value of \( \tau \) is small, the STDP learning window has a narrow width along the temporal direction. The STDP learning with such a narrow window has less chance for learning and the synaptic weights are not separated into maximum and minimum values completely. Because the synaptic weights of intermediate values change their values easily by LTP and LTD, the learning does not converge. The firing rates are highly influenced by the synaptic weights and exhibit sudden rise and fall. By the sudden change of the firing rates, the balance of synaptic weights breaks down significantly. Subsequently, the synaptic weights change with time by the STDP learning, which causes sudden change of firing rates again. By repeating these processes, ISO is reproduced.

6. Conclusion

In this paper, we investigated the occurrence of slow oscillation, or ISO, in a neural network with the STDP learning. We discovered that ISO is reproduced by the STDP learning with a narrowed window of the STDP function. With the narrow window, the learning does not converge and the synaptic weights become more influenced by the change of firing rates. From these results, it is indicated that ISO can be generated by the synaptic weights of intermediate values.

Acknowledgment

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. is supported by Grant-in-Aid for Research Activity Start-up (No. 26880020) from JSPS. The research of K.F. was supported by Grant-in-Aid for Challenging Exploratory Research (No 15K12137) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Challenging Exploratory Research (No. 24650116) from JSPS and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No.15KT0112) from JSPS.

References

Emergent oscillatory activities of plastic neural networks

Ryosuke Hosaka

Department of Applied Mathematics, Fukuoka University
Fukuoka 814-0180, Japan
Email: hosaka@fukuoka-u.ac.jp

Abstract—How simplify modeling properties are critical for results of the simulations. In neural network simulation, post-synaptic potentials are often modeled with Dirac delta function to reduce calculation costs. To investigate the effect of PSP modeling on resultant rhythmic activities of neural networks, we constructed two types of neural networks, class 1 and class 2 excitatory neurons, and it is developed with STDP learning rule. As a result, the stable rhythms were generated only in the case of the PSP modeling with delta functions. On other modeling cases, exponential and alpha functions, the rhythms were disappeared after the sufficient long learning periods. Therefore, using simplified PSP modeling should accompany careful handling to avoid erroneous simulation results.

1. Introduction

Neurons consists networks and are connected with each other via synapses. The synapses release neurotransmitters reacting to arrivals of action potentials. The neurotransmitters trigger post-synaptic potentials (PSP). The PSPs is modeled with alpha functions in realistic simulations. For simplicity, the PSP often modeled with exponential function. In most simplest cases, the Dirac delta function is used.

Ensemble activities of neurons, for instance, synchronous firing, cell-assembly formation, and rhythms of the network, are receiving great attention. They are seemed to play significant roles in the neural signal processing. We have reported a possible emergent mechanism of the neural synchrony [9]. In this paper, we rather focus on the rhythmic activities of neural networks.

The rhythmic activities are observed in the whole brain [12, 13, 14]. The categorization of the brain rhythms are based on their frequencies [2]: delta (1.5 ~ 4 [Hz]), theta (4 ~ 8 [Hz]), alpha (10 ~ 30 [Hz]), low gamma (30 ~ 80 [Hz]), and high gamma (80 ~ 200 [Hz]). In the hippocampus, the theta rhythms have a prominent role in coding of the animal’s position nesting high frequency oscillations [14]. In visual cortex, the gamma rhythms are related to the attention [5]. In motor cortex, the beta rhythms are dominant and increase during motor preparation [4].

Izhikevich demonstrated that a plastic spiking neural network can generate the delta and the gamma rhythms [11]. The neural network composed of 800 regularly spiking neurons for excitatory neurons and 200 fast spiking neurons for inhibitory neurons, and the neural network develops with STDP learning. In this work, the PSPs were modeled with the delta function. However, the alternative choice of PSP modeling function might lead to different consequences. Thus, we studied the effects of the PSP modeling on the rhythmic activities.

In mammalian neocortex, six fundamental classes of firing patterns are observed [3, 6, 7]: regularly spiking neurons; intrinsically bursting neurons; chattering neurons; fast spiking interneurons; low-threshold spiking neurons; and late spiking neurons. Among them, the regularly spiking neuron is most major neuron. Hodgkin stimulated the regularly spiking neurons by a constant current and observed its firing frequency [8]. By its excitability, he classified the regularly spiking neurons into two sub-categories: class 1 and class 2. The class 1 neurons start to fire with a low frequency through a critical point of firing. In contrast, the class 2 neurons start to fire with a high frequency that remains relatively constant even though the magnitude of the injected current increases. The class 1 and the class 2 excitabilities are realized by different bifurcation structures [15]: the class 1 excitability occurs when a neuron exhibits a saddle-node bifurcation; the class 2 excitability occurs when a neuron exhibits a Hopf bifurcation.

We constructed two types of neural networks with the class 1, the class 2 for excitatory neurons, and stimulate them by random inputs. The neurons are connected through chemical synapses, and the connection strength of synapses are dynamically changed depending on the activities of neurons. The dynamic change of synaptic connectivity is called Spike-Timing-Dependent synaptic Plasticity (STDP) [1]. To test if the choice of the PSP modeling significant effect on resultant rhythmic activities, the PSP of the class 1 or class 2 networks are modeled in delta, exponential or alpha functions. The results were compared in raster plots, power spectra, and the distributions of the plastic synaptic weights.

2. Methods

2.1. Post-synaptic potential

The PSPs were modeled in three ways: Dirac delta function, exponential function, and alpha function.
2.2. Neural network

In this paper, we used a neuron model proposed by Izhikevich [10] that is described as follows:

\[
\begin{align*}
\dot{v} &= 0.04v^2 + 5v + 140 - u + I(t), \\
\dot{u} &= a(bv - u),
\end{align*}
\]

with an auxiliary after-spike resetting condition:

\[
\text{if } v = 30 \text{[mV]}, \text{then } (v \leftarrow cu \leftarrow u + d).
\]

where \(v\) and \(u\) are dimensionless variables, \(a, b, c\) and \(d\) are dimensionless parameters, and \(\dot{}\) represents \(d/dt\), where \(t\) is the time (\([\text{ms}]\)). The variable \(v\) represents membrane potential (\([\text{mV}]\)) of the neuron and \(u\) represents a membrane recovery variable, which accounts for the activation of \(K^+\) ionic currents and inactivation of \(Na^+\) ionic currents, and it provides a negative feedback to \(v\).

We constructed neural networks in the following way. Each network is composed of 1,000 neurons, and 80% (or 20%) of the model neurons are excitatory (or inhibitory) as in the cortex. The first neural network has the class 1 excitatory neurons. The second neural network has the class 2 excitatory neurons. Properties of the inhibitory neurons are common for both networks. Excitable property of the inhibitory neuron is the class 2 and its time constant is much faster than the excitatory neurons as in the cortex.

We applied an STDP rule (details are described below) only to excitatory-to-excitatory connections while the other connections are fixed. Each neuron connected with only 100 other neurons. For simplicity, the time is assumed to be discrete (the time step is 1[ms]). Then, the dynamics of the neural networks develops as follows:

\[
\begin{align*}
\dot{v}_j(t + 1) &= v_j(t) + 0.04v_j^2(t) + 5v_j + 140 - u_j(t) + I_j(t) \\
&+ \sum_{i=1}^{N} w_{ij}(v_i(t - d_j) - 30), \\
\dot{u}_j(t + 1) &= u_j(t) + a_j(b_jv_j(t) - u_j(t) + e_j),
\end{align*}
\]

with the auxiliary after-spike resetting

\[
\text{if } v_j(t) = 30 \text{[mV]}, \text{then } (v_j(t) \leftarrow cu_j(t) \leftarrow u_j + d_j).
\]

where \(v_j(t)\) is membrane potential of the \(j\)-th neuron; \(u_j(t)\) is a recovery variable of the \(j\)-th neuron, and \(a_j, b_j, c_j, d_j\) and \(e_j\) are dimensionless parameters; \(e_j\) was introduced to regulate a firing rate of the neural network; For the class 1 excitatory neurons, \(a_j = 0.02, b_j = -0.1, c_j = -65.0, d_j = 8.0\) and \(e_j = -22\). For the class 2 excitatory neurons, \(a_j = 0.02, b_j = 0.26, c_j = -65.0, d_j = 8.0\) and \(e_j = 2\). For inhibitory neurons, \(a_j = 0.1, b_j = 0.2, c_j = -65.0, d_j = 2.0\) and \(e_j = 0\). \(w_{ij}\) is a synaptic connection from the \(i\)-th neuron to the \(j\)-th neuron. The synaptic weights from excitatory neurons are initially set to 6.0. The synaptic weights from inhibitory neurons are set to \(-5.0\). If the \(i\)-th neuron and the \(j\)-th neuron are not connected, \(w_{ij} = 0\). Self connection (\(w_{ii}\)) is also 0. \(h(t)\) is the PSP function (delta, exponential, or alpha functions). \(d_j\) is a synaptic transmission delay. The delay is decided randomly between 1 ~ 20 [ms]. \(I_j(t)=(0\ or\ 20)\) represents the external input for the \(j\)-th neuron, and \(I_j(t)\) follows a Poisson-process whose mean ISI is 1000 [ms].

2.3. STDP learning rule

Several experimental studies have reported window functions of the STDP learning (see e.g., Ref.[1]). In this paper, we used a typical function (Fig.7)[16]. The amount of synaptic weight modification (\(\Delta w\)) decreases exponentially with a temporal difference (\(\Delta t\)) between the arrival time of a pre-synaptic action potential, and the occurrence of its corresponding post-synaptic action potential:

\[
\Delta t = t_{pre} + d_{pre,post} - t_{post}
\]

where \(t_{pre}\) is spike time of a pre-synaptic neuron, \(t_{post}\) is spike time of a post-synaptic neuron, and \(d_{pre,post}\) is a delay time of spike transmission from the pre-synaptic neuron to the post-synaptic neuron. Then, synaptic modification \(\Delta w\) is described by the following equation,

\[
\Delta w(\Delta t) = \begin{cases} A_p e^{-\frac{\Delta t}{\tau_p}} & (\Delta t < 0), \\
-A_d e^{-\frac{\Delta t}{\tau_d}} & (\Delta t \geq 0),
\end{cases}
\]

where \(A_p\) and \(A_d\) are the maximum rate of modification (\(A_p = 0.1, A_d = 0.12\), \(\tau_p\) and \(\tau_d\) are the time constants for potentiation and depression, respectively (\(\tau_p = \tau_d = 20\ [\text{ms}]\)). We assumed that the synaptic efficacy is limited in the range of 0 \(\leq w_{ij} \leq 10\), because the STDP learning rule leads to further synaptic potentiation or depression to infinitely large or small synaptic weights.

3. Results

We first conducted the simulation with PSP modeling by the delta functions. Figure 1 shows raster plots of network activities. Dots on each raster plot indicate a firing of a neuron. In each raster plot, indices from 1 to 800 in vertical axis indicate the excitatory neurons, and the rests the inhibitory neurons. At the beginning of the simulations (in Fig.1 A), both the class 1 and the class 2 networks show slow rhythmic activities. These frequencies are 4 ~ 8 [Hz]. The slow rhythms correspond to the theta rhythm (4 ~ 8 [Hz]) that is often observed in hippocampus [14]. With time evolution, neurons become to fire in faster rhythms. The rhythm of the class 1 neural network settles down in beta frequency bands (Fig.1 B). The power spectrum of the rhythm is shown in Fig.1 C. The plastic synaptic weights are mostly biased to the lower bound (Fig.1 F). In contrast, the class 2 neural network generates the rhythms in high frequency bands at the end of the simulations (Fig.1 F). The frequency of the fast rhythm on the class 2 network corresponds to the gamma rhythm (30 ~ 80 [Hz], Fig.1 F). As the same with class 1 neural network, the plastic synaptic weights are biased to the lower bound (Fig.1 H).
Figure 1: Oscillatory activities of (A-D) class 1 and (E-H) class 2 neural networks with delta EPSP. (A-B) Raster plots at (A) 1st and (B) 60th seconds. (C) Power spectrum of oscillatory activities at 60th second. (E) The distribution of the plastic synaptic weights at 60th second. (E-H) The same as (A-D) but for class 2 neural networks.

Figure 2: The same as Fig.1 but for exponential EPSP.

Figure 3: The same as Fig.1 but for alpha EPSP.
We then conducted the simulation with PSP modeling by the exponential functions (Fig.2). At the beginning of the simulations (Fig.2, A and E), both the class 1 and class 2 networks showed faster oscillations than delta-shaped PSP function cases. With time evolution, neurons become to less fire. Eventually, both class 1 and 2 networks become asynchronous (Fig.2 B and F). We did not observe clear peaks in power spectra (Fig.2 C and G). The plastic synaptic weights are biased to the lower bound (Fig.2 D and H).

We finally conducted the simulation with PSP modeling by the alpha functions (Fig.3). At the beginning of the simulations (Fig.3, A and E), both the class 1 and class 2 networks showed slow oscillations like in delta-shaped PSP modeling case. However, the synchrony diminished soon, and kept asynchronous (Fig.3 B and F). We did not observe clear peaks in the power spectra (Fig.3 C and G). The plastic synaptic weights are biased to the lower bound (Fig.3 D and H).

4. Discussion

We constructed two neural networks, class 1 and class 2 excitatory neurons, and investigated the effect of PSP modeling on rhythmic activities. As a result, the rhythms are stable only for the delta PSP modeling. On the other modeling cases, the rhythms were gone after the sufficient learning.

The simple modeling is often employed in simulations, to achieve faster simulations and reduces calculation costs. Our results warn to careless use of PSP modeling. The choice of modeling functions have significant effect on the consequences. Namely, the stable rhythms were observed only in delta-function PSP modeling. For realistic simulations, the alpha function should be the first choice for the PSP modeling. To use more simplified PSP modeling should accompany careful handling to avoid erroneous simulations.

References


Connectivity analyses on self-organized neural competition networks

Hideyuki Kato†
†Tokyo University of Technology
1404-1 Katakura-machi, Hachioji-city, 192-0982, Tokyo, Japan
Email: katohdyk@stf.teu.ac.jp

Abstract—It was observed spike timing-dependent plasticity (STDP) induced competition among cells, that is called neural competition, when the network with axonal conduction delays exhibited oscillatory activity. Among such competitive cells, specific architecture might be constructed in the network even if uniform random connectivity was assumed under the initial condition. In the present study, it is demonstrated the STDP network organizes specific architecture by characterizing details of the network connectivity with the framework of the complex networks.

1. Introduction
Some theoretical studies reported STDP induced specific connectivity in recurrent networks [1–5]. In addition to these studies, it was shown a fraction of cells in a network obtained many strong connections (winners) and the others failed to do it (losers) by the interplay of three factors: oscillatory network activity, axonal conduction delays, and STDP [6]. This bias of in-coming and out-going connections might give alternative functional roles to the cells in the network. Based on this concept, in the present study, it is analyzed the connectivity among the cells in each population.

2. Materials and Methods
The network model, that was based on the simulator created by Izhikevich [7], used in simulations were same as in Ref. [6]. Since only connections between pairs of excitatory (E) cells were plastic, an E cell network was only focused on in this analysis. The cells were identified as winners (W) if its in-strength was larger than its out-strength and as losers (L) if vise versa (See Ref. [6]).

The network architecture was quantitatively evaluated with the synaptic strength matrix $W$. Additionally, the basic complex network statistics of the characteristic path length $L$ and the clustering coefficient $C$ were employed as the network quantifiers. There were two candidates of possible weight in the network in the network model: the axonal conduction delays and the synaptic strength. The analysis was, then, conducted with both types of weights. When the weight was the synaptic strength, the global efficiency $E$ was introduced instead of $L$ as a quantifier. The definition of $C$ in Ref. [9] was adopted since the evaluated networks were weighted and directional. For this analysis, connections with a positive value of synaptic weight were considered. The connectivity was evaluated from a network organized by STDP learning for 3,600 s.

3. Results
The following specific architecture emerged when neuron’s index was sorted by the descending order of the in-strength within each population. The connections of the W-cell population (WCP) almost died out and the extremely sparse subnetwork was organized. In contrast, the L-cell population (LCP) was relatively dense but surviving connections were on average weak. Remarkably, the strong pathway was established from the WCP to the LCP established, whereas the alternative direction disappeared.

In order to quantify the connectivity, $L$ (or $E$) and $C$ were calculated. In this analysis, the network evaluation was two steps. In the first step, by treating the STDP network as a digraph, the classical small-world test was conducted. For this, 1,000 random networks were generated from the STDP network by rewiring connections. In the second step, 1,000 weight-shuffled networks were generated from the STDP network by shuffling the weight values among connections.

In all the cases of the digraph and the weighted digraph analyses, the normalized $L$ (or $E$) and $C$ of the LCP were both unity, indicating the potentiation and the depression of the connections were spatially balanced within the LCP. In contrast to the LCP, $C = 1$ but $L$ was slightly larger than one in the digraph analysis of the WCP. This indicated spikes, on average, mediated via more cells to arrive at another cell.

When the axonal conduction delays were considered as the weight, $L = 1.2$ and $C = 1.05$. Then, this indicated, within the WCP, it took a long time for the spike transmission and three cells clusters were composed of relatively distant cells. For the weighted digraph, in which the weight was the synaptic strength, $E = 0.1$ and $C = 1$, indicating there were no pathway to reliably conduct spikes to the other cells and three cell clusters were not tightly linked.

4. Summery
In the present study, I characterized the network connectivity in the two populations organized by STDP. As a result, the winner cells organized a subnetwork that seemed to be inappropriate and inefficient for the information processing. It is unknown what functional roles exist in such a connectivity. Further analyses are needed to clarify this and this will be discussed anywhere else.

Acknowledgment
This work was supported by Grant-in-Aid for Young Scientists (B) (No.16K16127) from JSPS.

References
Computation of Visual Neurons in LGN–V1 Transmission

Kantaro Fujiwara† and Tohru Ikeguchi†

†Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6-3-1 Nijuku, Katsushika-ku, Tokyo, Japan
Email: kantaro@rs.tus.ac.jp

Abstract—In the lateral geniculate nucleus (LGN), it is reported that mean firing rates and higher order statistics change with time while a drifting sinusoidal grating was presented. On the other hand, such characteristics cannot be observed in the primary visual cortex (V1). Focusing on such differences, we construct a minimal cortical model of LGN–V1 transmission. As a result, we show that highly precise spike timings are essential for information coding in the LGN, while noisy spikes are utilized in V1.

1. Introduction

In the visual system, there exist a pathway from the lateral geniculate nucleus (LGN) to the primary visual cortex (V1), so-called the optic radiations. Neurons in the LGN send visual information coming from the retina to the V1. In the LGN of Macaca fascicular, it is reported that the mean firing rate changes with time while a drifting sinusoidal grating was presented [1]. In addition, we reported that temporally changing spike train local irregularity could be observed from the neurons in the LGN [2]. These facts indicate that neurons in the LGN use two types of neural coding scheme: firing rate coding and temporal coding. However, it is unknown how such complicated information treated in the LGN can be processed in the V1 area. In this study, we show the spike train local irregularity statistics observed from the V1 while a drifting sinusoidal grating was presented. A possible information processing role in LGN–V1 transmission is then discussed.

2. Response onset and offset in visual neurons

We applied spike train analysis methods in Ref. [3] to the spike data recorded from the LGN and V1 [1]. The data we used are publicly available from the Neural Signal Archive ([4], http://www.neuralsignal.org). The recordings were obtained while a drifting sinusoidal grating was presented. From the results of temporally changing behavior of the spike train statistics observed from the LGN, irregularity of the spike train is high in the early stage, then gradually decreases, then finally increases again. These changes were independent of the mean firing rates. On the other hand, in V1 area, there was no significant change of the statistics independent of firing rates.

For reproducing characteristic behaviors of the statistics in V1, we construct a minimal spiking neuron model. By introducing large amplitude of noise for V1 neuronal input, characteristic V1 statistics can be observed. On the other hand, highly precise spike timings with much smaller values of noise are essential for reproducing characteristic statistics in the LGN.

3. Conclusion

In this paper, we used spike train analysis methods proposed in Ref. [3] to physiological spike data [4], and found different coding schemes between the LGN and V1. By introducing noisy LGN–V1 pathway, characteristic behaviors of the statistics in V1 can be reproduced. These results in the LGN and V1 may provide novel insights to neural codes in visual information processing.

Acknowledgments

The research of K. F. was supported by Grant-in-Aid for Challenging Exploratory Research (No.15K12137) and by Grant-in-Aid for Young Scientists (B) (No.16K16138) from JSPS. The research of T. I. was partially supported by Grant-in-Aid for Exploratory Research (No.24650116) and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No.15KT0112) from JSPS.

References

A Computational Model for Pitch Pattern Perception with the Echo State Network

Miwa Fukino†, Yuichi Katori‡,§, and Kazuyuki Aihara¶

†Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
‡School of Systems Information Science, Future University Hakodate
116-2 Kamedanakano-cho, Hakodate, Hokkaido 041-8655, Japan
§Institute of Industrial Science, The University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
Email: fukino@sat.t.u-tokyo.ac.jp, katori@fun.ac.jp, and aihara@sat.t.u-tokyo.ac.jp

Abstract—The predictive coding theory assumes that the sensory system of the cortex continuously predicts incoming stimuli and detects residual errors. The mismatch negativity (MMN) is a neural response to a deviation of learned regularity, and is regarded as an error signal in this theory. Here we report a preliminary study on a computational model of the auditory MMN using the Echo State Network which is one of the recurrence neural network models. We trained the network by an oddball task with two pitch patterns. The result shows that our model simulates a qualitatively similar waveforms with the MMN response to deviant pitches.

1. Introduction

The auditory mismatch negativity (MMN) (for review, see [1]) is a neural response to a deviation of learned regularity. It is one of the event related potentials which arises between 100 and 200 ms after the deviant stimulus onset, and can be measured by electroencephalography (EEG) or magnetoencephalogram (MEG). When a deviant magnitude is higher, the MMN amplitude is larger and the latency is smaller. When a deviant probability is lower, the MMN amplitude is larger and there is no effect for the latency.

The MMN has been widely used in clinical and theoretical studies, for example, as a biomarker of psychosis [2, 3], and as a research of brain plasticity in terms of comparing the difference between musicians’ and non-musicians’ MMN [4, 5]. However, neurophysiological mechanisms of the MMN are still controversial.

The MMN is often regarded as an error signal of the predictive coding model [6] in recent studies [7]. Several mathematical models of the MMN are proposed based on the predictive coding idea (e.g. [8, 9]). Wacongne et al. (2012) proposed a neuronal model of the auditory cortex accounting for the MMN, whose scheme is as follows. There are four components in the model: Thalamic inputs, Prediction error layer, Predictive layer, and Memory trace. An input sound stream is composed of two pitches A and B. Population of neurons in the Predictive layer continuously predicts incoming Thalamic inputs. Population of neurons in the Prediction error layer receives two inputs: inhibitory inputs coming from the Prediction layer and excitatory inputs, or the Thalamic inputs, and then the residual is the error signal, namely MMN. The error signal is transmitted to the Prediction layer to adjust the internal model of the prediction. Memory neurons keep the inputs of past few hundred milliseconds.

Wacongne et al. used the spiking neuron model [10], and implemented precise neuronal behavior in terms of realistic receptors (AMPA, NMDA, and GABA), synapses, and spiking neurons. However, it is difficult to apply this model for processing more complex sound patterns, because it focuses on the precise neuronal behavior of the primary auditory cortex, and it does not consider more complex auditory patterns.

The Echo State Network (ESN) [11, 12, 13] is one of artificial recurrent neural networks, where neurons are sparsely and randomly connected, and only outputs are trained. The scheme of ESN is as follows. $x_i(n)$ is the $i$th neuron ($i = 1, \ldots, N$) at time $n$ in the dynamical reservoir. $d_i(n)$ and $y_j(n)$ are the $i$th teacher data and system output ($j = 1, \ldots, L$) at time $n$, respectively. $W^{rec}$, $W^{back}$, and $W^{out}$ are weight matrices of recurrent connections inside the dynamical reservoir, feedback connections from outputs to reservoir, and output connections from the reservoir to system outputs, respectively.

The activation of internal units is updated according to

$$x(n + 1) = f(W^{rec}x(n) + W^{back}y(n)),$$

(1)

$$y(n + 1) = f^{out}(W^{out}(x(n + 1), y(n)),$$

(2)

where $f$ and $f^{out}$ denotes the individual units’ transfer functions. The internal weights $W^{rec}$ and feedback weights $W^{back}$ are set to random and sparse at first, and not changed during the training. Only the output weights $W^{out}$ are trained. The echo state network can learn nonlinear systems, although the structure is simple.
In the present study, we propose the predictive coding model with ESN for processing complex and realistic auditory patterns, and provide a learning procedure for the proposed network model.

2. Model

We propose a computational model for pitch pattern perception based on Jaeger’s echo state network model [11] and the predictive coding architecture of Wacongne et al.’s neuronal model [9]. Structure of the proposed model is shown in Fig. 1.

![Figure 1: Structure of proposed model](image)

2.1. Overall structure of the network

\( x(n) \) is a status of memory trace layer at time \( n \) (\( N_x \) dimensions), \( y(n) \) is a status of predictive layer (\( N_y \) dimensions), \( r(n) \) is a status of prediction error layer (\( N_r \) dimensions), and \( d(n) \) is a status of given sensory inputs (\( N_d \) dimensions).

The state of each layer is updated according to

\[
\begin{align*}
    x(n+1) &= x(n) + \frac{\delta}{\tau}(-\alpha_0 x(n) + f_r(W_{rec} x(n - k_x)) \nonumber \\
    &+ W_{back}(y(n - k_y) + r(n - k_r))), \\
    y(n) &= f_y(W_{out} x(n)), \quad \text{and} \quad \text{r}(n) = f_r(d(n) - y(n)),
\end{align*}
\]

where \( W_{rec} \) is an \( N_x \times N_y \) weight matrices of recurrent connections, \( W_{back} \) is an \( N_y \times N_r \) weight matrices of feedback connections, and \( k_x, k_y, k_r \) are delay times. We define \( f_r(x) = \tanh x, f_y(x) = \tanh x, \) and \( f_x(x) = \max(x, 0) \). Initial state of \( x(n) \) is randomized as a uniform distribution over \(-0.2 \sim 0.2\).

2.2. Configuration of Network

**Step1**: Setting \( W_{rec} \) and \( W_{back} \)

We define the network of proposed model based on [11].

1. Generate an internal weight matrix \( W_0 \). Assign 1 or \(-1\) to randomly selected \( \beta_0 N_r \times N_y \) components of \( W_0 \). Assign 0 to the other units.

2. Normalize \( W_0 \) to a matrix \( W_1 \) with unit spectral radius by putting \( W_1 = 1/|\lambda_{max}| W_0 \), where \( \lambda_{max} \) is maximum eigenvalue of \( W_0 \).

3. Scale \( W_1 \) to \( W_{rec} = \alpha_r W_1 \).

4. For \( W_{back} \), first, randomly assign \(-1 \sim 1\) to \( \beta \times N_y \times N_r \) components of \( W_2 \) with a uniform distribution. Assign 0 to the other units. Then, normalize the sum of columns of the unit vectors. Assign \( \alpha_r \) and set it to \( W_{back} \).

**Step2**: Training \( W_{out} \)

In our proposed method, the weight values for the output connection \( W_{out} \) are computed by a few iterative epochs.

1. **As initial values of \( W_{out,0} \), all the components of \( W_{out,0} \) are set to 0.**

2. **Using \( W_{out,(m-1)} \) which is calculated in the \( m \)-th epoch, the \( m \)-th weight matrix \( W_{out,m} \) is calculated.** Using the sensory input \( d(n) \) with the time ranges \( n = 0, ..., T_1 \), the model is driven with “teacher-focusing”, which means that the feedback of output state \( y(n) \) is replaced with the teacher signal (the given sensory input) \( d(n) \) as following equations.

\[
\begin{align*}
    x(n+1) &= x(n) + \frac{\delta}{\tau}(-\alpha_0 x(n) + f_r(W_{rec} x(n - k_x)) \nonumber \\
    &+ W_{back}(y(n - k_y) + r(n - k_r))) + \sigma \xi(n), \quad (6) \\
    y(n) &= f_y(W_{out,m-1} x(n)), \quad (7) \\
    r(n) &= f_r(d(n) - y(n)), \quad (8)
\end{align*}
\]

where \( \xi(n) \) has random values of normal distribution with mean 0 and variance 1. \( \sigma \) specifies the amplitude of random values. \( x(n) \) and \( y(n) \) at time from \( T_0 \) to \( T_1 \) are used to calculate \( W_{out,m} \). Next, input the time series of \( x(n) \) into state collecting matrix \( M \), where \( M \) is \((T_1 - T_0 - 1) \times N_x \) matrix. Then, input sigmoid-inverted \( d(n) \) or \( \tanh^{-1} d(n) \) into \( G \), where \( G \) is \((T_1 - T_0 - 1) \times N_x \) matrix. \( W_{out,m} \) is calculated using ridge regression as

\[
(W_{out,m})^T = (M^T M + \lambda E)^{-1} M^T G, \quad (9)
\]

where \( \lambda \) is a coefficient for adjusting a sparseness, and \( E \) is a unit matrix.
3. After the $I$ iterations of above calculation, we get $W_{out}^{new,I}$ as $W_{out}$ by repeating the learning for several times ($I = 10$ iterations are enough).

3. Results

We use an input data simulating the oddball paradigm which is often used for the stimulus of MMN experiments. The red curves in Fig.2 (b) shows the input data. The lower and upper curve indicate pitch A and B, respectively. Repetition of two pitches A and B make a stream of patterns: the data includes 80% of standard patterns “AAAB” and 20% of deviant pitches (e.g., “AAAA”).

We demonstrate our model with the input data simulating the oddball paradigm using following parameter values. $N_c = 200$, $N_t = 2$, $\alpha_c = 0.6$, $\beta_c = 0.1$, $\alpha_d = 0.8$, $\beta_d = 0.1$, $\alpha_0 = 0.7$, $\tau = 2.5$, $\delta = 1$, $k_r = 5$, $k_c = 10$, $k_y = 0$, $\sigma = 10^{-2}$, and $\lambda = 0.1$.

Figure 2 shows the typical response of the proposed model. The red and blue curves of (b) are the teacher data ($d(n)$ on the Sensory input of Fig.1) and resulting data ($y(n)$ on the Predictive layer) respectively. Many red curves of (a) are time series of all neurons on the dynamical reservoir ($x(n)$ on the Memory trace layer). Green curves of (c) are time series of residual between the teacher and resulting data ($r(n)$ on the Prediction error layer).

The response on the dynamical reservoir (many red curves of Fig.2 (a)) vary among units, and a set of the whole units represent the resulting data (blue curves in Fig.2 (b)).

The time series $r(n) = f_r(d - y)$ (green curves in Fig.2 (c)) is the residual between the the Sensory input (red curves of Fig.2 (b)) and the prediction, or the resulting data (blue curves in Fig.2 (b)). They are correspond to the error signal of predictive coding, or the MMN. We can observe that responses on the predictive layer (the blue curves of Fig.2 (b)) are reproducing the sensory input (red curve). Especially, they are almost similar at the timing of peaks, that is, A or B sound is played. On the contrary, when the sounds are omitted or deviant, the response on the prediction layer shows wrong predictions. Note that the deviant sound (AAABAAABAAAAB) is played at the timing of a black arrow. There is a small predicted wave arising at the upper blue curve.

Thus, our model simulates qualitatively reproduce waveform of the MMN response to deviant pitches.

This report is the first step towards a computational model of the MMN. It is an important future problem to analyze this model by changing parameters, and discuss the relation between the computational and implementation levels of analysis in terms of Marr’s levels of analysis.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 16K00246. This work was funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).

References


Encoding Multi-Dimensional Time Series Data with Reservoir Computing

Yuichi Katori

The School of Systems Information Science, Future University Hakodate, 116–2 Kamedanakano-cho, Hakodate, Hokkaido, Japan, Institute of Industrial Science, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo, Japan.

Email: katori@fun.ac.jp

Abstract—Encoding of given time series data can be basis of various information processing and a key to understanding properties of the dynamical system that generate the time series. Efficient encoding requires to extracts features of the data and decompose the given data with the features. Features on time series can be represented as a set of kernel functions, and a given data can be decomposed using the kernel functions and sparsification of the representation. This sparse and shiftable kernel method efficiently encodes one-dimensional time series data into a series of point on time as a point process. Reservoir computing paradigm provides a strategy to model multi-dimensional time series data with randomly coupled non-linear elements. Here we propose to combine the shiftable kernel method and the reservoir computing paradigm. We use the echo state network, one of implementation of the reservoir computing, and propose that the reservoir computing can be utilized to dynamically organize the kernel functions and to efficiently encode multi-dimensional time series. We demonstrate that a complex multi-dimensional time series data can be encoded into a few points in the point process with the proposed method.

1. Introduction

Dynamical system modeling has played a crucial role in the development of techniques for data processing of many research fields. For instance, nonstationary acoustic structures of timing relations among acoustic events or harmonic periodicities provide cues for many types of auditory processing, e.g. sound localization and speech recognition. Time series data reflecting activities of the brain or a local neural circuit provide information underlying neural processing and are applicable for brain-machine interface. These time series data reflect properties of underlying complex dynamical system and may include redundancy and exhibit repetition of a certain pattern of time course. Reducing the redundancy, extracting features, and encoding the data are key to understand the underlying dynamical systems and may provide a basis of applications for pattern recognition, regression, and prediction.

Temporal features on time series can be characterized to a precise point in time, such as the onset of sound or a task cue. Signal decomposition using a linear superposition of time-shiftable kernel function [1, 2] works well to extract features of the data and to encode the data into point process. This approach works well particularly on sound data recorded as a one-dimensional time series.

Most of the above-mentioned time series data are recorded with multiple sensors, e.g. electroencephalogram, multi-unit recording of spike trains, and imaging data containing many pixels, and microphone array. Even sound data recorded as one-dimensional time series are often analyzed as multi-dimensional time series of the spectrogram. Each channel of these multi-dimensional time series may have temporal and spatial correlations with each other reflecting the structure of an underlying dynamical system.

The key of encoding such multi-dimensional data is to reduce redundancy among the channels. In one view, efficient coding theory, the goal of the modeling is to encode the maximal amount of information about the input signal by statistically independent features. Current methods like principal component analysis or independent component analysis perform this processing with an assumed spatial distributions of the data. However, there is no efficient coding strategy based on both spatial and temporal feature of multi-dimensional time series data.

Reservoir computing is a paradigm of understanding and training recurrent neural networks[3], where neurons are sparsely and randomly connected. In the reservoir computing, supervised adaptation of all weight value of recurrent connection is not necessary, and only training a memoryless supervised readout from the dynamical reservoir (recurrent network) is enough to obtain flexible and multiple time courses. One form of the reservoir computing paradigm is the echo state network[4, 5], which composed of simple non-linear elements.

Here we propose that the reservoir computing paradigm can be utilized to encoding multi-dimensional time series and that kernel functions for extracting features of a given time series data can be generated with dynamics of the echo state network. In this paper, we explain the generative form of the model and algorithm for the encoding, then show preliminary results.
2. Methods

Here we use the echo state network [4, 5] and the sparse and shiftable kernel method of signal representation [1, 2]. The basic idea of our approach is to dynamically organize a set of kernel functions with the echo state network and to efficiently represent the correlated multi-dimensional time series data with this set of kernel functions.

2.1. Generative Model with Reservoir Computing

In the proposed model (Figure 1), a given \( N_d \)-dimensional signal \( d(n) \) is represented by the output of the dynamical reservoir \( y(n) \). State of the dynamical reservoir \( x(n) \) and output \( y(n) \) are updated according to following equations.

\[
\begin{align*}
x(n + 1) &= x(n) + \frac{\delta}{T} - \alpha_0 x(n) \\
&\quad + f_i \left( W^{rec} x(n) + W^{back} y(n) + W^{in} u(n) \right), \\
y(n) &= f_i( W^{out} x(n) ),
\end{align*}
\]

where \( W^{rec} \) is \( N_x \times N_r \) weight matrices of recurrent connections on the dynamical reservoir, \( W^{back} \) is \( N_r \times N_r \) weight matrices of feedback connections from the output layer to the dynamical reservoir, \( W^{in} \) is \( N_r \times N_u \) weight matrices of input connections on from the input layer to the dynamical reservoir. We define \( f_i(x) = \tanh x \). \( \delta = 1, \tau \) specifies the time scale of the dynamical reservoir.

The dynamical reservoir is driven by the input \( u(n) \), which is given as an ensemble of a kernel function \( \phi \). Here we use Gaussian function as the kernel \( \phi(n) = \exp(-n^2/\sigma^2) \) with the width of kernel \( \sigma \), and \( k \) th component of \( u(n) \) is represented as follows.

\[
u_k = \sum_i s_k^i \phi(n - \tau^k_i),
\]

where \( \tau^k_i \) and \( s_k^i \) are the temporal position and coefficient of the \( i \) th instance of \( k \) th component of \( u(n) \). \( n_k \) is the number of instance on the \( k \) th component of \( u(n) \).

2.2. Encoding Algorithms

Above described equations specify the generative form of the model but does not provide an encoding algorithm. The optimal values of \( \tau^k_i \), \( s_k^i \), and \( W^{out} \) for a given signal are necessary to be computed with randomly generated \( W^{rec} \), \( W^{back} \), and \( W^{in} \).

As to the first step for generating \( W^{rec} \), \( W^{back} \), and \( W^{in} \), we follows the procedures in [5]. For \( W^{rec} \), first, randomly generate \( N_r \times N_r \) matrix \( W_0 \) by assigning 1 or -1 to randomly selected \( \beta_0 N_r \times N_r \) components of \( W_0 \) and assign 0 to the others. Then, normalize \( W_0 \) to a matrix \( W_1 \) with unit spectral radius by putting \( W_1 = 1/|\lambda_{max}|W_0 \), where \( \lambda_{max} \) is maximum eigenvalue of \( W_0 \). Finally, scale \( W_1 \) to \( W^{rec} = \alpha_1 W_1 \). For \( W^{back} \), first, randomly assign -1 \sim 1 to \( \beta \times N_r \times N_r \) components of \( W_0 \) with a uniform distribution and assign 0 to the other units. Then, normalize the sum of columns of the unit vectors, multiply with coefficient \( \alpha_2 \), and set it to \( W^{back} \). \( W^{in} \) is generated by same manner with coefficient \( \alpha_3 \).

In the second step, optimal values of \( \tau^k_i \), \( s_k^i \), and \( W^{out} \) are computed with a given time series data. The objective of this optimization is to minimize the error while maximizing coding efficiency. Here we use the following equation as the objective function.

\[
E = \sum_{i} \sum_{n} (y_i(n) - d_i(n))^2 + \lambda \sum_{k} \sum_{n} |u_k(n)|
\]

The first term on the r.h.s of equation is the error and the second term is a term for enhancing sparseness of the input.

We try to minimize \( E \) by finding optimal list of instances \( (\tau^k_i, s_k^i) \) with iterations described below. In each iteration step, find the optimal \( \tau^k_i \) and \( s_k^i \) with brute force approach, namely adopt \( (\tau^k_i, s_k^i) = \arg \min_{\tau^k_i, s_k^i} E \) as an instance on the input layer.

In each iteration step, \( W^{out} \) are computed by driving the model with “teacher-focusing”, which means that the feedback of output state \( y(n) \) is replaced with the teacher signal (the given input) \( d(n) \) as following equations.

\[
\begin{align*}
x(n + 1) &= x(n) + \frac{\delta}{T} - \alpha_0 x(n) \\
&\quad + f_i \left( W^{rec} x(n) + W^{back} d(n) + W^{in} u(n) \right), \\
x(n) \quad \text{and} \quad y(n) \quad \text{in time from} \quad T_0 \quad \text{to} \quad T_1 \quad \text{are used to calculate}
\end{align*}
\]
W_{\text{out}}$. Align the time series of $x(n)$ into state collecting matrix $M$, where $M$ is $(T_1 - T_0 - 1) \times N_x$ matrix. Then, input sigmoid-inverted $d(n)$, or $\tanh^{-1}(d(n))$ into $G$, where $G$ is $(T_1 - T_0 - 1) \times N_x$ matrix. $W_{\text{out}}$ is calculated using pseudo-inverse of $M$ as

$$
(W_{\text{out}})^T = M^{-1}G.
$$

Continue the iteration till the $E$ does not decrease, and the set of instances of the input layer is obtained.

3. Results

We demonstrate our methods with an artificially generated multi-dimensional time series (Figure 2). This time series is generated as an ensemble of spatially and temporally distributed Gaussian functions on $N_x = 20$ dimension of given data (See red curves in Fig. 2(c)).

We use the dynamical reservoir of size $N_x = 100$ and $N_y = 5$ input layer with following parameter values. $\alpha_r = 0.2, \beta_r = 0.1, \alpha_b = 0.2, \beta_b = 0.1, \alpha_i = 1, \beta_i = 0.1, \tau = 2.5, \sigma = 5$, and $\lambda = 1.0$. For simplicity, we set $s_i^2 = 1$ for this simulation.

Figure 2 shows that the given multi-dimensional time series is encoded with three instances on the input layer, which virtually corresponds to three points of marked point process (Fig. 2(a)). The input drives the dynamical reservoir and causes specific pattern of fluctuations (Fig. 2(b)). This fluctuation is read out to the output layer so that the output layer reproduces the given time series data (Fig. 2(c)). The responses on the output layer well reproducing the given time series data (Fig. 2(d)).

The proposed method uses random numbers for generating weight matrix in the first step of the encoding, and thus the performance of the encoding depend on the realization of the weight matrix, whereas, the performance is stable for this specific time series data, and result shown in Fig. 2 is a typical case.

4. Discussion

Here we propose a new encoding method for multi-dimensional time series data based on reservoir computing and sparse and shiftable kernel method. Our preliminary result shows that complex multi-dimensional time series can be virtually encoded into a few points in the point process.

The present study may provide insights for understanding efficient coding mechanisms of biological systems, i.e. auditory system, particularly, the higher order representation of auditory coding and animal vocalizations.

In the future, we should evaluate details of the proposed method, namely, the dependency of the performances on the many parameters and network structure. To apply this method to large-scale and realistic problem, more efficient ways of configuring connection matrix and encoding algorithms are necessary. Furthermore, we should apply this method to real engineering problems, e.g. encoding audio signals, communication, analysis of biomedical data including applications for brain-machine interfaces.

Acknowledgments

The author would like to thank NOLTA2016 organizing committee members for their fruitful suggestions and comments. This work was supported by JSPS KAKENHI Grant Number 16K00246.

References


Reproduction of nonlinear cochlea response by asynchronous bifurcation processor

Kentaro Takeda† and Hiroyuki Torikai

†Department of Computer Science, Kyoto Sangyo University
1–3 Motoyama, Kamigamo, Kitaku, Kyoto 605-8555, Japan
Email: g1344793@cse.kyoto-su.ac.jp, torikai@cse.kyoto-su.ac.jp

Abstract—Biological cochleae have highly nonlinear responses to sound stimuli, e.g., response curves (so-called tuning curves) are highly nonlinear with respect to stimulation frequencies. In this paper, the nonlinear dynamics of a Hopf-bifurcation-type cochlea model based on a concept of an asynchronous bifurcation processor is investigated. It is shown that the model can reproduce nonlinear tuning curves of not only a mammalian cochlea (cat) but also a reptilian cochlea (turtle).

1. Introduction

Biological cochleae have highly nonlinear responses to sound stimuli such as nonlinear frequency tuning curve, multi-tone suppression, first and second pitch shifts, adaptation, parallel spike density encoding, and so on [1]. Among such nonlinear responses, the nonlinear frequency tuning curve is focused in this paper. Fig. 1 shows an example of the frequency tuning curve of a reptilian cochlea (turtle). In this figure, a frequency at a minimum peak of the curve is called a characteristics frequency and is corresponding to a specific position of a basilar membrane in the cochlea, i.e., the basilar membrane works as a kind of nonlinear mechanical Fourier transformer. There are many mathematical and circuit models for the cochlea [2]-[10]. One of simple-but-powerful mathematical cochlea models is the Hopf-cochlea [6]-[10]. On the other hand, in this paper, the following four discrete states \{X, Y, P, Q\} are used in a cochlea model.

\[ X \in \mathbb{Z}_N = \{0, \cdots, N-1\}, \quad Y \in \mathbb{Z}_N, \]
\[ P \in \mathbb{Z}_M = \{0, \cdots, M-1\}, \quad Q \in \mathbb{Z}_M, \]

where \(N\) and \(M\) are positive integers, which determine the resolution of the state space \{(X, Y, P, Q) | X \in \mathbb{Z}_N, Y \in \mathbb{Z}_N, P \in \mathbb{Z}_M, Q \in \mathbb{Z}_M\}.

Based on the normal form of the Hopf-bifurcation, cochlea models have been presented and investigated [8]-[10]. On the other hand, in this paper, the following four discrete states \{(X, Y, P, Q) | X \in \mathbb{Z}_N, Y \in \mathbb{Z}_N, P \in \mathbb{Z}_M, Q \in \mathbb{Z}_M\} are used in a cochlea model.

\[ X \in \mathbb{Z}_N = \{0, \cdots, N-1\}, \quad Y \in \mathbb{Z}_N, \]
\[ P \in \mathbb{Z}_M = \{0, \cdots, M-1\}, \quad Q \in \mathbb{Z}_M, \]

where \(N\) and \(M\) are positive integers, which determine the resolution of the state space \{(X, Y, P, Q) | X \in \mathbb{Z}_N, Y \in \mathbb{Z}_N, P \in \mathbb{Z}_M, Q \in \mathbb{Z}_M\}.

2. Cochlea Model based on Asynchronous Bifurcation Processor

Let

\[ t \in \mathbb{R} \]

be a continuous time and let

\[ x_1 \in \mathbb{R}, \quad x_2 \in \mathbb{R} \]

be continuous states. Also, let \(\alpha \in \mathbb{R}\) be a continuous parameter. Then the following set of equations are known as the normal form of the Hopf-bifurcation [14].

\[ \frac{dx_1}{dt} = \alpha x_1 - x_2 + x_1(x_1^2 + x_2^2), \]
\[ \frac{dx_2}{dt} = x_1 + \alpha x_2 + x_2(x_1^2 + x_2^2). \]

Based on the normal form of the Hopf-bifurcation, cochlea models have been presented and investigated [8]-[10]. On the other hand, in this paper, the following four discrete states \{(X, Y, P, Q) | X \in \mathbb{Z}_N, Y \in \mathbb{Z}_N, P \in \mathbb{Z}_M, Q \in \mathbb{Z}_M\} are used in a cochlea model.

\[ x_1 \in \mathbb{R}, \quad x_2 \in \mathbb{R} \]

be continuous states. Also, let \(\alpha \in \mathbb{R}\) be a continuous parameter. Then the following set of equations are known as the normal form of the Hopf-bifurcation [14].

\[ \frac{dx_1}{dt} = \alpha x_1 - x_2 + x_1(x_1^2 + x_2^2), \]
\[ \frac{dx_2}{dt} = x_1 + \alpha x_2 + x_2(x_1^2 + x_2^2). \]

Based on the normal form of the Hopf-bifurcation, cochlea models have been presented and investigated [8]-[10]. On the other hand, in this paper, the following four discrete states \{(X, Y, P, Q) | X \in \mathbb{Z}_N, Y \in \mathbb{Z}_N, P \in \mathbb{Z}_M, Q \in \mathbb{Z}_M\} are used in a cochlea model.

\[ X \in \mathbb{Z}_N = \{0, \cdots, N-1\}, \quad Y \in \mathbb{Z}_N, \]
\[ P \in \mathbb{Z}_M = \{0, \cdots, M-1\}, \quad Q \in \mathbb{Z}_M, \]

where \(N\) and \(M\) are positive integers, which determine the resolution of the state space \{(X, Y, P, Q) | X \in \mathbb{Z}_N, Y \in \mathbb{Z}_N, P \in \mathbb{Z}_M, Q \in \mathbb{Z}_M\}.

In this paper, a cochlea model based on the ABP and the Hopf-cochlea is investigated. It is newly shown that the model can reproduce nonlinear tuning curves of a reptilian cochlea (turtle). Note that our previous work suggests the model can reproduce a mammalian cochlea (cat) [10], and thus this paper suggests that the model can be used to model cochleae of many species.

Figure 1: Rough sketch of frequency tuning curves of a reptilian (turtle) cochlea [13].
The discrete states \( X \) and \( Y \) are used to reproduce nonlinear oscillatory behaviors of the basilar membrane, and the discrete states \( P \) and \( Q \) are used to control velocities of the discrete states \( X \) and \( Y \), respectively. Then the following functions \( g_x : Z_N \times Z_N \rightarrow \mathbb{R} \) and \( g_y : Z_N \times Z_N \rightarrow \mathbb{R} \) are introduced.

\[
g_x(X,Y) = \begin{cases} 
M - 1 & \text{if } \text{Int} \left( \frac{-1}{g_x(X,Y)} \right) \geq M - 1, \\
-(M - 1) & \text{if } \text{Int} \left( \frac{-1}{g_x(X,Y)} \right) \geq -(M - 1), \\
\text{otherwise} & 
\end{cases}
\]

\[
g_y(X,Y) = \begin{cases} 
M - 1 & \text{if } \text{Int} \left( \frac{-1}{g_y(X,Y)} \right) \geq M - 1, \\
-(M - 1) & \text{if } \text{Int} \left( \frac{-1}{g_y(X,Y)} \right) \geq -(M - 1), \\
\text{otherwise} & 
\end{cases}
\]

where \( T_X \in \mathbb{R}^+ \) and \( T_Y \in \mathbb{R}^+ \) are periods of the following internal clocks \( C_X(t) \) and \( C_Y(t) \), respectively.

\[
C_X(t) = \begin{cases} 
1 & \text{if } t = 0, T_X, 2T_X, \cdots, \\
0 & \text{otherwise}, 
\end{cases}
\]

\[
C_Y(t) = \begin{cases} 
1 & \text{if } t = 0, T_Y, 2T_Y, \cdots, \\
0 & \text{otherwise}. 
\end{cases}
\]

Note that the ratio \( T_X/T_Y \) of the periods can be either rational (i.e., the clocks \( C_X \) and \( C_Y \) are phase-locked or synchronized) or irrational (i.e., the internal clocks \( C_X \) and \( C_Y \) are asynchronous). Since the clock generators are uncoupled, the period ratio \( T_X/T_Y \) is generically irrational and thus the internal clocks \( C_X(t) \) and \( C_Y(t) \) are generically asynchronous. Recall that the asynchronicity of the internal clocks \( C_X(t) \) and \( C_Y(t) \) is used to realize a smooth vector field of our model. The internal clocks \( C_X \) and \( C_Y \) trigger the following transitions of the discrete states \( P \) and \( Q \), respectively.

If \( C_X(t) = 1 \), then

\[
P(t+) := \begin{cases} 
P(t) + 1 & \text{if } P(t) < |F_X|, \\
0 & \text{if } P(t) \geq |F_X|. 
\end{cases}
\]

If \( C_Y(t) = 1 \), then

\[
Q(t+) := \begin{cases} 
Q(t) + 1 & \text{if } Q(t) < |F_Y|, \\
0 & \text{if } Q(t) \geq |F_Y|. 
\end{cases}
\]

The symbol \( " \lim_{t \to +t} + \epsilon \)" denotes an "instantaneous state transition" hereafter. The internal clocks \( C_X \) and \( C_Y \) trigger the following transitions of the discrete states \( X \) and \( Y \), respectively.

If \( C_X(t) = 1 \) and \( P(t) \geq |F_X| \), then

\[
X(t+) := \begin{cases} 
X(t) + 1 & \text{if } X(t) \neq N - 1 \text{ and } F_X \geq 0, \\
X(t) - 1 & \text{if } X(t) \geq 0 \text{ and } F_X < 0, \\
X(t) & \text{otherwise}. 
\end{cases}
\]

If \( C_Y(t) = 1 \) and \( Q(t) \geq |F_Y| \), then

\[
Y(t+) := \begin{cases} 
Y(t) + 1 & \text{if } Y(t) \neq N - 1 \text{ and } F_Y \geq 0, \\
Y(t) - 1 & \text{if } Y(t) \geq 0 \text{ and } F_Y < 0, \\
Y(t) & \text{otherwise}. 
\end{cases}
\]

Let \( s(t) = A \sin(2\pi ft) \) be a stimulation signal. Then the following stimulation input \( S(t) \) is introduced.

\[
S(t) = \begin{cases} 
1 & \text{if } t = t_1, t_2, t_3, \cdots, \\
-1 & \text{if } t = t_n, t_{n+1}, t_{n+2}, \cdots, \\
0 & \text{otherwise}, 
\end{cases}
\]

where the instantaneous density of the pulse positions \( \{t_1, t_2, t_3, \cdots, \} \) is proportional to \( s(t) \) for \( s(t) > 0 \) and is zero for \( s(t) \leq 0 \); and the instantaneous density of the pulse positions \( \{t_n, t_{n+1}, t_{n+2}, \cdots, \} \) is proportional to \( -s(t) \) for \( s(t) < 0 \) and is zero for \( s(t) \geq 0 \). Note that such a density modulation can be easily realized by using a standard density modulator. The stimulation input \( S(t) \) triggers the following transitions of the discrete state \( P \) and \( Q \).

If \( S(t) = 1 \) and \( F_X \geq 0 \), then

\[
P(t+) := \begin{cases} 
P(t) + 1 & \text{if } P(t) < |F_X|, \\
0 & \text{if } P(t) \geq |F_X|. 
\end{cases}
\]

If \( S(t) = -1 \) and \( F_X \geq 0 \), then

\[
P(t+) := \begin{cases} 
P(t) - 1 & \text{if } P(t) > 0, \\
F_X & \text{if } P(t) = 0, 
\end{cases}
\]

If \( S(t) = 1 \) and \( F_X < 0 \), then

\[
P(t+) := \begin{cases} 
P(t) - 1 & \text{if } P(t) > 0, \\
F_X & \text{if } P(t) = 0, 
\end{cases}
\]
Then, the stimulation input $S(t)$ triggers the following transitions of the discrete state $X$ and $Y$.

If $S(t) = 1$ and $F_X \geq 0$, then

$$X(t_+) := \begin{cases} X(t) + 1 & \text{if } X(t) \neq N - 1 \\ X(t) & \text{otherwise} \end{cases}$$

and $P(t) = 0$, otherwise.

If $S(t) = -1$ and $F_X \geq 0$, then

$$X(t_+) := \begin{cases} X(t) - 1 & \text{if } X(t) \neq 0 \\ X(t) & \text{otherwise} \end{cases}$$

and $P(t) = 0$, otherwise.

If $S(t) = 1$ and $F_X < 0$, then

$$X(t_+) := \begin{cases} X(t) + 1 & \text{if } X(t) \neq N - 1 \\ X(t) & \text{otherwise} \end{cases}$$

and $P(t) \geq \lvert F_X \rvert$, otherwise.

If $S(t) = -1$ and $F_X < 0$, then

$$X(t_+) := \begin{cases} X(t) - 1 & \text{if } X(t) \neq 0 \\ X(t) & \text{otherwise} \end{cases}$$

and $P(t) \geq \lvert F_X \rvert$, otherwise.

Fig. 2 shows a typical time chart of the model and Fig. 3 shows typical time waveforms of the model. In order to characterize the time waveforms, the following $RMS$ of the discrete state $X$ is introduced.

$$RMS(X) = \lim_{T \to \infty} \frac{1}{T} \int_{-T/2}^{T/2} (X(t) - N/2)^2 dt.$$
Figure 4: Frequency tuning curves of our model, where parameter values are the same as those in Fig. 3. The dashed line shows audible range which have 30-700[Hz] of a turtle [13]. It can be seen that the curves mimic some of the physiologically measured frequency tuning curves of the turtle in Fig. 1.

The characteristics curve of the minimum sound power level $SPL_{min}$ with respect to the input frequency $f$ is said to be a frequency tuning curve.

Fig. 4 shows frequency tuning curves of our model, where a dashed line shows audible range 30-700[Hz] of a turtle [13]. Comparing the frequency tuning curves in Fig. 4 (our model) and that in Fig. 1 (turtle), it can be seen that our model can mimic the physiologically measured frequency tuning curves of the turtle.

3. Conclusions

The nonlinear dynamics of a Hopf-bifurcation-type cochlea model based on a concept of an asynchronous bifurcation processor was investigated. It was shown that the model can reproduce nonlinear tuning curves of not only a mammalian cochlea (cat) but also a reptilian cochlea (turtle). Future problems are including (a) analysis of the occurrence mechanism of the bifurcation by using discrete-continuous hybrid Poincare map, (b) comparison of a frequency tuning curve of an FPGA-implemented circuit of our model with biological data, (c) reproductions of more highly nonlinear responses to sound stimuli such as multitone suppression, first and second pitch shifts, adaptation, parallel spike density encoding, and so on. This work was partially supported by JSPS KAKENHI Grant Number 15K00352.

References

Network analyses of chaotic systems

Miki U. Kobayashi† and Yoshitaka Saiki‡

†Faculty of Economics, Rissho University, 4-2-16 Osaki, Shinagawa-ku, Tokyo 141-8602, Japan
‡Graduate School of Commerce and Management, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan
Email: miki@ris.ac.jp

Abstract—It is found that chaotic dynamical systems possess infinitely many unstable periodic orbits embedded in a chaotic set. A chaotic motion can be considered as an irregular transition process among the unstable periodic orbits. This irregular transition process is better described as a network. Recently, we have proposed a concrete method for describing a chaotic motion as a network, in which the unstable periodic orbits are regarded as nodes and links are provided based on the transition process of a chaotic motion among the unstable periodic orbits.

General networks can be roughly divided into the following three kinds; regular, random and complex networks. Various real networks are complex networks, i.e. power and neural networks, networks in linguistics, and world wide web and so on. It is known that smaller average path length, larger clustering coefficient and scale free degree distribution are typical properties of complex networks. A problem for kinds of networks of chaos is open.

We consider the Lorenz system: \( \dot{x} = -10x + 10y, \quad \dot{y} = -xz + rx - y, \quad \dot{z} = xy - (8/3)z \). Networks of the Lorenz system are constructed by our scheme. Average path length, clustering coefficient and degree distribution are analysed in the network structures of the Lorenz system. We reveal that networks of Lorenz chaos is a kind of complex networks. Furthermore, we discuss network structures of the system with \( r = 28 \) whose the system is singular hyperbolic and with \( r = 60 \) whose the system is non-hyperbolic.

References

A hierarchical routing algorithm for MANET based on multi-agent learning

Yuki Hoshino†, Hidehiro Nakano‡ and Arata Miyauchi‡

†Tokyo City University, 1-28-1, Tamazutsumi, Setagaya-ku, Tokyo, 158-8557, Japan
Email: g1581521@tcu.ac.jp
‡Proceedings of the 2nd International Conference on Future Information Technology, Japan, November 27th-30th, 2016

Abstract—Mobile Ad-hoc Networks (MANETs) can construct impromptu networks by wireless mobile nodes without fixed infrastructure. A routing algorithm for MANETs based on multi-agent cost learning has been proposed. In this algorithm, multiple agents learn the costs of the communication paths by calculating weighted values, and select the effective paths based on the costs. This algorithm is robust for the movements of wireless mobile nodes. In order to develop this algorithm into larger-scale environments, this paper proposes a hierarchical routing algorithm for MANETs. The proposed algorithm constructs hierarchical networks based on address information, and does not require cluster-head nodes. In the simulation experiments, the effectiveness of the proposed algorithm can be confirmed by evaluating the delivery rate and control packet counts.

1. Introduction

Mobile Ad-hoc Networks (MANETs) can construct impromptu networks by wireless mobile nodes without fixed infrastructure[1]. Therefore, routing protocol for MANETs requires adaptability for dynamic network topology by node mobility. Types of routing Algorithm for MANETs are classified into "Proactive" type that each node sends routing information at fixed time intervals, and "Reactive" type that each path is searched just before each node sends data packets. "Proactive" type is suitable for high node density networks and high frequency communications. "Reactive" type is suitable for high mobility networks.

AntHocNet[2] has a property of "Proactive" type and "Reactive" type. A routing algorithm for MANETs based on multi-agent cost learning[3] is more effective than AntHocNet in dynamic networks. But, this method can use one routing information to one destination. And, it can’t use routing information when it failed to search the path.

Therefore, this paper proposes a hierarchical routing algorithm for MANETs. The proposed method can use routing information to destination node even when it failed to search the path, by using hierarchical routing tables. In the simulation experiments, the effectiveness of the proposed algorithm can be confirmed by evaluating the delivery rate and control packet counts.

2. Routing algorithm for MANETs based on multi-agent cost learning

A routing algorithm for MANETs based on multi-agent cost learning has been proposed[3]. In this algorithm, multiple agents learn the cost of the communication paths, in order not to select the same intermediate nodes, and search unknown paths based on the costs. In this algorithm, the $i$th node has a routing table $T_i$, whose example is shown in Table 1. In this table, $W$ denotes a weighted value learned by calculating costs between nodes. For example, $W_{ij}$ is the weighted value in the case where the $i$th node sends data packets to the destination node $d_1$ via the $j$th neighbor node. The $i$th node sends hello message for the neighbor node management. As the $i$th node detects a neighbor node $j$, the weighted value $W_{ij}$ is registered to the table $T_i$. If the other nodes have been already registered in $T_i$, the value of $W_{ij}$ is set to the average value of $W$ in $T_i$. Otherwise, the value of $W_{ij}$ is set 1 as the initial value.

By the following three stages, each node sends data packets to the destination node.

1. Forward Agent
2. Backward Agent

2.1. Forward Agent

Let us consider that a source node sends data packets to a destination node. First, the source node sends plural forward agent by broadcast transmission. As each intermediate node that receives one of forward agents sends it to a neighbor node selected by Eq. (1).

$$P_{ij} = \frac{W_{ij}^{\alpha}}{\sum_{n \in N} W_{in}^{\alpha}}$$

where $P_{ij}$ is probability that $i$th node selects the $j$th node as the next hop node. $\alpha$ is priority for weights. The $i$th node...
updates $W_{ij}$ by Eq.(2) after the $i$ th node sent the forward agent.

$$W_{ij} \leftarrow W_{ij} + Cost_{ij}$$

(2)

If an intermediate node that received forward agent has a destination node in the neighbor nodes, the intermediate sends to the destination node as the next hop node. The forward agent constructs an address list of all nodes which it passes through. In addition, only a forward agent firstly arriving at the destination node is accepted, and the other forward agents are rejected.

2.2. Backward Agent

When a forward agent firstly arrived at the destination node, the destination node sends backward agent. A backward agent follows the address list constructed by the forward agent in reverse. An intermediate node that receives the backward agent sends it to the next hop node based on the address list. Then, the intermediate node updates the weighted value $W_{ij}$.

$$W_{ij} \leftarrow W_{ij} \times (1 - \rho)$$

(3)

where $\rho$ is a reduction rate.

2.3. Data Packet Forwarding

The source node that receives backward agent sends data packets. Each intermediate node to the destination node selects the next hop node by Eq.(4).

$$s = \begin{cases} 
\text{Eq.}(1) & \text{if } x \leq \epsilon_0 \\
\arg \max_{n \in N_i} W_{in}^{-\alpha} & \text{otherwise} 
\end{cases}$$

(4)

where $\epsilon$ is the random value in the range of $[0, 1]$. $\epsilon_0$ is the threshold of the probability in the range of $0 \leq \epsilon_0 \leq 1$, and decides relative importance of exploitation and exploration. Then, the data packets are delivered to the destination node.

3. Proposed Method

The proposed method brings a hierarchical property to the conventional method. The conventional method requires that at least one forward agent arrives to a destination node in order to learn the path. Also, the source node and the intermediate nodes can’t reuse routing information for the other destination nodes. The proposed method appends hierarchical routing to the conventional method to solve the problems.

The proposed method has two routing network layers, high network layers and a low network layer. Each high network layer has plural nodes whose address consists of the same lower bits. The low network layer sends data packets between high layer nodes. The proposed method uses the following three routing tables.

- Destination group table
- High level layer table
- Low level layer table

The destination group table is used in routing when a neighbor node belonging to the same group as the destination node is searched. In the high level layer, there are the nodes belonging to the same group. The high level layer table is used in routing when the next hop node in the high level layer is searched. The low level layer table is used in routing to the next high level layer node. The high level layer nodes are classified by the forward agent and backward agent.

3.1. Routing

Figure 1 shows an example of routing. The source node has address 0x65 and the destination node has address 0xFA. Lower 4 bits are used as the group indexes. That is, the group indexes of the source node and the destination node are 5 and A, respectively.

At first, by using the destination group table, the source node (0x65) sends a forward agent to a neighbor node belonging to the same group as the destination node. In Fig.1, A is the destination group, and the node having address 0xEA is the sub-destination. By using the low level layer table, the forward agent is sent from the source node to the sub-destination node.

As the forward agent arrived at the sub-destination node, by using the high level layer table, it is sent from the sub-destination node to the next hop node in the high level layer node. In Fig.1, the node having address 0xEA selects the node having address 0x1A as the next hop node in the high level layer. By using the low level layer table, the forward agent is sent from the sub-destination node. Repeating in this manner, the forward agent arrives at the destination node.

Next, the backward agent is sent from the destination node to the source node. the backward agent follows the address list by the forward agent to reverse and also updates the routing table. As the backward agent is sent to the next node, the unused address list entry is removed. In this moment, if flag in the current address list entry is unset, a node update "low level layer table". If it is set, a
node updates "high level layer table". If there are no flag in all address list entries, a node updates "Destination group table".

### 3.2. Collecting same group addresses

In order to realize the high level layer routing, it is required to grasp the high level layer network topology from the address of each node. Each node collects the address information of the nodes belonging to the same group from the address lists of forward agents and backward agents. If the unknown address is found from an address list, the address is collected as one of the high level layer nodes. Each a backward agent is sent from a destination node via some intermediate nodes, and have the address and hop counts of these nodes. Then, the address information and the hop counts of these nodes are registered in the routing table. Collecting the information, each node grasps the high level layer network topology, and constructs each routing table.

### 3.3. Sub backward agent

From a source node, plural forward agents are sent to a destination node. One of them firstly arriving at the destination node changes into a backward agent. In the conventional method, the paths through which the other forward agents pass are not learned. Therefore, in the proposed method, if a forward agent passes through a path between high level layer nodes and the corresponding backward agent does not pass through the path, a sub backward agent is sent between the high level layer nodes and learns the path. The timing to send the sub backward agent is decided by the elapsed time when the forward agent starts from the high level layer node.

### 4. Simulation Experiments

We examine the performances of the proposed method on the following environments, and perform the experiments, by comparing with the conventional method [3].

#### 4.1. Experiment 1

At first, we perform the experiments for the environment in which all nodes are fixed. Table 2 shows the parameters in the experiments. In this experiments, transmission delay and transmission data size is not considered. For each 0.25 second, a randomly selected source node sends data packets to a randomly selected destination node. We measure delivery rate rate and control packet counts. The lower 3 bits in the address of each node is used as group indexes.

Figures 2-5 show the results for delivery rate and control packet counts with different area size.

As shown in Figs. 2 and 4, the delivery rate of the proposed method more quickly increases than that of conventional method. On the other hand, as shown in Figs. 3 and 5, control packet counts of the proposed method is more than those of the conventional method.

#### 4.2. Experiment 2

Next, we perform the experiments for the environment in which all nodes move. Each node moves according to Random Way Point Model [4]. Pause time is 30 sec and max speed is 5 m/sec. The other parameters for the environment are shown in Table 2. In the experiments, area size is fixed to 2400m × 800m.

Figures 6 and 7 show the delivery rate and the control packet counts, respectively.

As shown in Fig. 6, even if each node moves, the delivery rate of the proposed method more quickly increases.
than that of the conventional method. As the area size becomes larger, this tendency becomes more conspicuous.

Delivery rate of the proposed method more quickly increases than that of the conventional method. And, the control packet counts of the proposed method is more than those of the conventional method. Because, the proposed method can reuse routing information for the other destination node by using the high level layer. Also, the search range can be reduced by using the hierarchical routing, compared with the case of searching all nodes in the network. Those result show that the proposed method is more effective on large scale environments. The proposed method can more immediately communicate on large networks.

5. Conclusion

The proposed method introduces a hierarchical property to the conventional method. According to it, delivery rate more quickly increases than that of the conventional method. On the other hand delivery rate of the conventional method increases very slow on large environments. Therefore, the proposed method is more effective than the conventional to the environments.

Control packet counts of the proposed method are more than those of conventional method. Therefore, we reduce packet counts on the proposed method as future works.

References


An improved routing method using transmission memory information for wireless communication networks

Tokuma Sato†, Takayuki Kimura‡ and Kenya Jin’no‡

†Graduate School of Electronics, Information and Media Engineering, Nippon Institute of Technology
‡Department of Electrical and Electronic Engineering, Nippon Institute of Technology
4-1-1 Gakuendai, Miyashiro, Minami-Saitama, 345-8501 Japan
Email: 2158009@estu.nit.ac.jp, tkimura@nit.ac.jp, jinno@nit.ac.jp

Abstract—Recently, a drastic increase in mobile users leads to packet congestion on the wireless communication network. Many researchers then try to propose efficient routing methods that remove the packet congestion even if the flowing packets increase. As one of the effective routing methods for the wireless communication networks, an effective routing method using transmission memory information has already been proposed. However, the method used fixed parameters at each node to route the packet. In this study, we improve the routing method [1] using dynamic parameter adjustment method to apply our method to various topologies of the computer networks.

1. Introduction

Traffic control of the computer networks is one of the most important problems for reliable communication. Therefore, many researchers propose various routing methods for communication networks. For example, the routing method using distance between nodes and transmission memory information in wired communication networks has been proposed in Ref. [2]. Yang et al.[3] proposed the routing method using geographical distance between nodes and the number of holding packets. This approach improves the packet congestion by using the information of congestion status of neighbor nodes. The effectiveness of routing method using transmission memory information for the wireless communication network model has been proposed in Ref. [1]. The routing method [1] alleviates the packet congestion using the diversification of selecting paths of packets which is realized by memory information. However, this method used the fixed parameter at each node. These parameter settings make it harder to apply the routing method to various topologies of the networks. To overcome this problem, we introduce a dynamic parameter adjustment method to the routing method [1] in this study. From the results of the numerical experiments, the improved routing method using transmission memory information transmits the packets more quickly than the conventional routing method [1].

2. Mobile communication network model

We use the weighted and undirected graphs $G = (V,E)$ to construct the mobile communication network model, where $V$ is the set of nodes and $E$ is the set of links. Each node represents a wireless agent and each link represents a connection between two nodes. $N = |V|$ expresses the total number of nodes and these $N$ nodes are placed in the square-shaped cell of size $L$. The $i$th node updates its position using the following equations:

$$x_i(t+1) = x_i(t) + v \cos \theta_i(t), \quad (1)$$
$$y_i(t+1) = y_i(t) + v \sin \theta_i(t), \quad (2)$$
$$\theta_i(t) = \phi_i, \quad (3)$$

where $x_i(t)$ and $y_i(t)$ are coordinates of the $i$th node at the $t$th time. In addition, we define $v$ is the moving speed of the node and $\phi_i (i = 1, \ldots, N)$ is the phase position of the $i$th node. $\phi_i$ is randomly determined using the uniformly distributed random numbers from $-\pi$ to $\pi$. If the $i$th node exceeds the range of the square-shaped cell size $L$, the position of the $i$th node is updated again not to move beyond the area of network. In addition, the distance between the $i$th node and the $j$th node at the $t$th iteration is defined as follows:

$$D_{ij}(t) = \sqrt{(x_i(t) - x_j(t))^2 + (y_i(t) - y_j(t))^2}. \quad (4)$$

If the distance of two nodes (the $i$th node and the $j$th node) is smaller than communication radius, $C_r$, the $j$th node becomes neighbor node of the $i$th node or connected to the $i$th node. Then, the node selects one of the neighbor nodes within its communication range as a transmitting node. The packets are transmitted to the next node based on the First-In First-Out basis; a packet at the head of the buffer of the node is transmitted to its neighbor node, and the packet is store
at the tail of the neighbor node. Sources and destinations of packets are randomly selected using uniformly distributed random numbers. If the buffer of a transmitting node is full, the transmission of the packet to node is canceled. These packets wait for the next opportunity for transmission.

3. Packet routing using transmission memory information

In this study, if the distance between the $i$th and the $j$th nodes at the $t$th iteration is less than or equal to its communication range, namely $D_{ij}(t) \leq C_r$, we set the distance between these nodes to 1. On the other hand, we set the distance of the nodes to $\infty$. Each node determines the transmitting node using following equations.

\[
y_{ij}(t+1) = \xi_{ij}(t+1) + \zeta_{ij}(t+1),
\]

where $\xi_{ij}(t+1)$ is the distance information and $\zeta_{ij}(t+1)$ is the transmission memory information of the $i$th node at the $t$th iteration. The distance information is defined as follows:

\[
\xi_{ij}(t+1) = \frac{d_{ij}(p_i(t))(t)}{C_r}, j \in N_i(t),
\]

where $p_i(t)$ is a transmitting packet from the $i$th node at the $t$th time, $g(p_i(t))$ is destination of $p_i(t)$, $N_i(t)$ is the number of the neighbor nodes of the $i$th node at the $t$th time, $d_{ij}(p_i(t))(t)$ is the distance between the $j$th neighbor node and $g(p_i(t))$.

The transmission memory information of the $i$th node is defined as follows:

\[
\zeta_{ij}(t+1) = \alpha \sum_{\gamma=0}^{t} k_r^\gamma x_{ij}(t-\gamma)
\]

\[
= \alpha x_{ij}(t) + k_r \zeta_{ij}(t-1),
\]

where $\alpha$ is parameter that determines the strength of transmission memory information, and $k_r$ is decay parameter. $x_{ij}(t)$ is the packet transmission history of the $j$th node at the $t$th time and defined by,

\[
x_{ij}(t) = \begin{cases} 1 & (\min(y_{ij}(t+1)), \\ 0 & (\text{otherwise}) \end{cases}
\]

In Eq. (8), the $i$th node transmits a packet to the $j$th neighbor node if $y_{ij}(t+1)$ is the smallest among the neighbor nodes.

We first evaluate the performance of conventional method. These experiments are conducted as follows. First, packets are generated in the network whose sources and destinations are randomly assigned. Then, the density of the packets $D = \rho \cdot N \cdot B$, is set to between $0 < \rho \leq 1$, where $\rho$ is the packet generating rate, $N$ is the number of nodes, and $B$ is the buffer size of each node. The square-shaped cell size, $L$, is set to 4. The number of nodes, $N$, is set to 100 and the buffer size of each node, $B$, is set to 100. In addition, the moving speed of node, $v$, is set to 0.01 and the communication radius, $C_r$, to 0.6. The parameters in Eqs. (7) are set as follows: $\alpha = 6.0$ and $k_r = 0.9$.

The number of iterations $I$, is set to 3,000. Here, we defined a single iteration as the determinations of the transmitting nodes and the transmissions of packets to those nodes.

![Figure 1: Results by the conventional method for the mobile communication networks](image)
Figure 1 shows the total number of arriving packets, an average number of hops of arriving packets, and an average times of the arriving packets by the conventional method. In Fig. 1(a), the number of arriving packets by $\alpha = 1$ increases in comparison to those by $\alpha = 10$ if the density of packet $D$ is less than 0.5. On the other hand, the number of arriving packets increases if $\alpha$ is set to 10 as $D$ becomes large. In Figs. 1(b) and (c), the average number of hops and the average number of times increase as $\alpha$ becomes large. From these results, the performance of the conventional method is varied depending on the state of the networks: a small $\alpha$ setting is better for lower $D$ case and a high $\alpha$ setting is better for high $D$ case. Then, if each node can automatically adjust the value of $\alpha$ depending on the state of the network, we expect that the performance of our routing method will be enhanced.

4. A proposed method with improved transmission memory information

From the previous results, we confirmed that the routing method might be enhanced if the parameter $\alpha$ automatically changes depending on the state of the networks. To realize this functionality, we change the transmission memory information as follows:

$$\zeta_{ij}(t + 1) = \beta n_i(t) \sum_{\gamma=0}^{t} k_{ij}^\gamma e_{ij}(t - \gamma),$$

$$= \beta n_i(t) x_{ij}(t) + k_{ij} \zeta_{ij}(t - 1),$$

where $n_i(t)$ is the number of holding packets of the $i$th node, $\beta$ is a controlling parameter.

We compared the performance of the proposed with that of the conventional method. We apply the same experimental conditions used in the previous experiments. In addition, the parameter $\alpha$ in Eq. (7) is set to 6.0 and $\beta$ in Eq. (9) is set to 0.05.

Figure 2 shows the number of arriving packets, the average number of arriving packets, and the average times of arriving packets by the conventional method and the proposed method. In Fig. 2(a), the number of arriving packets by the proposed method is larger than that by the conventional method. In addition, in Fig. 2(b), the average hops of the arriving packets by the proposed method is smaller than that by the conventional method. Further, in Fig. 2(c), the packets by the proposed method are transmitted to their destinations more quickly than the conventional method. From the results of the numerical experiments, the proposed method transmits the packets by effectively adjusting the strength of the transmission memory in comparison to the conventional routing method.

5. Conclusions

In this study, the performance of the routing method using transmission memory information has been evaluated for the mobile communication networks. The conventional method used fixed parameter for every node to route packets. We then improved the routing method by using the dynamic adjusting method for determining the strength of transmission memory at each node. As a result, we confirmed that the performance of the proposed method is enhanced in comparison to
the conventional method.

Evaluation of our method using the realistic network will be conducted for the future works.

The research of T.K. was partially supported by a Grant-in-Aid for Young Scientists (B) from JSPS (No. 16K21327).

References


A Study of a probabilistic routing method based on packet transmission capacity

Sho Tagami†, Takayuki Kimura‡ and Kenya Jin’no‡

†Graduate School of Electronics, Information and Media Engineering, Nippon Institute of Technology
‡Department of Electrical and Electronic Engineering, Nippon Institute of Technology
4-1-1 Gakuendai, Miyasiro, Minami-Saitama, 345-8501 Japan
Email: e1112308@estu.nit.ac.jp, tkimura@nit.ac.jp, jinno@nit.ac.jp

Abstract—Recently, the data packet flow increases on the Internet because of wide spread of mobile devices or high-speed wireless communications. The data packets are often lost or delayed if the communication networks are congested. The shortest path protocol that is commonly used in the real-world system is not effective for routing the packets because the one transmits data packets using only the distance information of the communication networks. Therefore, the packet congestion easily occurs at the hub nodes. Thus, an efficient routing method which effectively controls the data flow in the communication network is desired to remove the packet congestion. From this view point, a probabilistic routing method for the communication networks has been proposed. In this method, each node decides the transmitting nodes using probabilistic way based on the number of holding packets and degree of the adjacent nodes. However, this method doesn’t consider the transmission capacity. Thus, the packets are transmitted to the nodes that store large amount of packets exceeding the transmission capacity. The packets are then accumulated at these nodes, and this causes the congestion on the network. From this view point, we proposed another probabilistic packet routing method using the number of holding packets at adjacent nodes and transmission capacity in this study. From the results of numerical experiments, the proposed method effectively avoids the congestion of packet in comparison to the conventional shortest path routing and the conventional probabilistic routing method.

1. Introduction

Recently, the data packet flow increases on the Internet because of wide spread of mobile devices or high-speed wireless communications. The data packets are often delayed because of the congestion in the communication networks. As the solution for tackling these problems, using access restriction is implemented. However, it doesn’t consider the users satisfaction.

It has been already reported that the communication networks have scale-free property[1, 2]. Then, the shortest path protocol[3] that is commonly used in the real-world system is not effective for routing the packets because the one transmits data packets using only the distance information of the communication networks. As a result, the packet congestion easily occurs at the hub nodes. Thus, an efficient routing method which effectively controls the data packets flow in the communication networks is desired to remove the packet congestion[4, 5]. From this view point, a probabilistic routing method for the communication networks has been proposed[6]. In this method, transmitting node of a packet is determined using the number of holding packets and degree of adjacent nodes. Although this routing method can alleviate the packet congestion, we confirmed that the packets are still trapped at the nodes. The reason why the packets are trapped at some nodes is that the transmission capacity is evaluated used for routing packets by the probabilistic method. From this view point, we proposed a new probabilistic packet routing method using the number of holding packets and transmission capacity at adjacent nodes. From the results of numerical experiments, the proposed method effectively avoids the congestion of packet in comparison to the conventional routing methods.

2. Network model

In this study, we use a simple undirected graph composed by nodes and links as a communication network model. We also employ the scale-free networks[7] as a topology of communication network. The scale-free network is generated by the following steps. First, a complete graph in which every node has degree of $K$ is generated. Then, a node which has degree of $k$ is added to an existing network. Each branch of newly added node is connected to the nodes using preferential selection probability $\Pi(k_i) = k_i / \sum_{j=1}^{N} k_j$, where $k_i$ is the degree of the $i$th node, $N$ is the number of nodes in the existing network. Node attachment is repeated until a terminating condition, $N = \bar{N}$, is satisfied.

Figure 1 shows an example of scale-free networks by $N = 100$ and $K = 3$. In Fig. 1, the nodes which have large circle stand for the hub nodes. Figure 2 shows the degree distribution of scale-free networks. In Fig. 2, although a few nodes have the large degree, most of the nodes have low degree.

In this study, each node has a buffer to store the packets and the packets are transmitted to the adjacent nodes using First-In First-Out rule. Moreover, a single iteration is defined as determinations of the transmitting nodes at the nodes and transmissions of packets. The $i$th node transmits...
3. A probabilistic routing method using degree and holding packet information

A conventional probabilistic routing method[6] uses the degree information and the number of holding packets at adjacent nodes to transmit the packets to their destinations. The conventional method determines a transmitting node using the following probability:

\[
P_i(t) = \frac{k_i(n_i(t) + 1)^\alpha}{\sum_j k_j(n_j(t) + 1)^\alpha}.
\]

In Eq. (1), \(P_i(t)\) is a probability that the \(i\)th node is selected as the transmitting node at the \(t\)th iteration. \(k_i\) is the degree of the \(i\)th node, \(n_i(t)\) is the number of holding packets in the \(i\)th node at the \(t\)th iteration, \(\alpha\) is a controlling parameter. If a destination of a packet is found among adjacent nodes, the packet is transmitted to the destination without using Eq. (1). By using Eq. (1), although a packet is easily transmitted to the hub nodes if \(\alpha\) takes large value, the packets are hardly transmitted to the hub nodes if \(\alpha\) takes negative value.

Figure 1: An example of scale-free network (\(N = 100\)).

Figure 2: Degree distribution of scale-free networks.

\[p(k) = \frac{1}{k} \frac{1}{\Gamma(\alpha+1)} \left(\frac{k}{\lambda}\right)^{\alpha} e^{-\frac{k}{\lambda}}\]

\[p(k) = \frac{1}{k} \frac{1}{\Gamma(\alpha+1)} \left(\frac{k}{\lambda}\right)^{\alpha} e^{-\frac{k}{\lambda}}\]

Figure 3: Conventional probabilistic routing method[6]

Figure 3 shows schematic diagrams for each parameter. In Fig 3, the transmission probability of the \(i\)th node and \(i_2\)th node are denoted by \(P_{i_1}\) and \(P_{i_2}\). The transmission probability then changes by the value of \(\alpha\) in Eq. (1). In Fig. 3(a), the transmitting probability to the \(i_1\)th node which has large degree and the large number of holding packets increases if \(\alpha\) is positive. On the other hand, if \(\alpha\) is negative, the transmitting probability to the \(i_2\)th node increases. In this study, packet congestion rates for various \(\alpha\) cases are used to evaluate the performance of conventional probabilistic routing method. The packet congestion rate is defined as follows:

\[
\eta = \frac{1}{R \Delta t} \frac{\sum_{t=1}^{T-\Delta t} \Delta N p}{(T - \Delta t)}.
\]
In Eq. (2), \( R \) is the number of packets generated at each iteration. \( \Delta N_\pi \) is defined by \( \Delta N_\pi = N(t + \Delta t) - N(t) \) where \(< \cdots >\) indicating an average over the time windows of width \( \Delta t \), and \( N(t) \) is the number of packets in the network at the \( r \)th iteration. The packet congestion rate \((0 \leq \eta \leq 1)\) indicates that the network is congested as \( \eta \) approaches 1. In the numerical simulations, we set the number of nodes \( N = 100 \) and the number of degree \( K = 3 \). In addition, we repeated the packet transmission for 1000 iterations. Then, we conducted the numerical experiments for 10 times and the averaged the results.

In Eq. (3), \( P_i(t) \) is the probability that the \( i \)th node is the transmitting node for routing the packets using the following probability:

\[
P_i(t) = \sum_j \frac{(n_i(t) + 1)^\gamma}{C_j}.
\]

In Eq. (3), \( P_i(t) \) is the probability that the \( i \)th node is selected as the transmitting node at the \( r \)th time, \( C_i \) is the transmission capacity of the \( i \)th node, \( n_i(t) \) is the number of holding packets in the \( i \)th node at the \( r \)th time, \( \gamma \) is a controlling parameter. In the same way as the conventional probabilistic routing method, if the node finds the destination of a packet among the adjacent nodes, the packet is delivered to its destination without using Eq. (3). By using Eq. (3), the packets are easily transmitted to the nodes which have the packets less than their transmission capacities if \( \gamma \) takes negative value. We compared the congestion rate by the conventional probabilistic method[6] with that by proposed method. Same experimental conditions used in the previous section are applied to these numerical simulations. The parameters in Eq. (1) and Eq. (3) are set as follows: \( \alpha = -2.0 \) and \( \gamma = -10.0 \).

Figure 5 shows the average number of packets at each node for the conventional probabilistic routing method. In Fig. 5, the numbers of packets of nodes with high degree are larger than their transmission capacities. The congestion is then accumulated at these nodes because these nodes cannot transmit the packets over their transmission capacities. This problem seems to lie in the fact that the transmission capacity of the adjacent node is not considered to route the packet. We then propose a new probabilistic packet routing method that incorporates transmission capacity and the number of holding packets at the nodes.

4. A probabilistic routing method using transmission capacity and the number of holding packets

In the previous section, we clarify that the congestion spreads to whole network because the transmission capacity is ignored to route the packets. We then propose a new probabilistic routing method. In our proposed method, each node determines the transmitting node for routing the packets using the following probability:

\[
P_i(t) = \sum_j \frac{(n_i(t) + 1)^\gamma}{C_j}.
\]
5. Conclusion

In this study, we proposed a probabilistic routing method which incorporates not only the number of holding packets but also transmitting capacity for routing the packets. First, we confirmed that the conventional routing method cannot remove the packets at the hub nodes because the conventional method doesn’t consider the transmission capacities. From this viewpoint, we proposed the probabilistic routing method using transmission capacity and the number of holding packets. From the results of numerical experiments, the proposed method avoided the congestion of packets by using the transmitting capacity effectively.

In future works, we apply our proposed method to various topologies of realistic communication networks.

The research of T.K. was partially supported by a Grant-in-Aid for Young Scientists (B) from JSPS (No.16K21327).

References

A solution strategy for packet routing problem by chaotic neurodynamics with degree information

Yuuki Morita†, Takayuki Kimura‡ and Kenya Jin’no‡

†Graduate School of Electronics, Information and Media Engineering, Nippon Institute of Technology
4–1 Gakuendai, Miyashiro, Minami-Saitama, Saitama, 345-8501, Japan
‡Department of Electrical and Electronics Engineering, Nippon Institute of Technology
4–1 Gakuendai, Miyashiro, Minami-Saitama, Saitama, 345-8501, Japan
Email: 2168014@estu.nit.ac.jp, tkimura@nit.ac.jp, jinno@nit.ac.jp

Abstract—Recently, because of the rapid spread of mobile devices, the number of flowing packets in the computer network increases, and this causes the congestion of the packets. As one of the effective packet routing methods to remove the packet congestion, a packet routing method using chaotic neurodynamics has already been proposed. By using the chaotic neurodynamics that realizes decentralized routings of packets, this method shows higher arrival rate of the packets than the shortest path routing. However, if the number of packets drastically increases, the method shows poor performance because packets are congested at the hub nodes. To remove the congestion at hub nodes, we proposed a new routing strategy using chaotic neurodynamics with degree information in this study. From results of the numerical simulations, the proposed routing method alleviates the packet congestion at the hub nodes effectively, achieving large number of packets are transmitted to their destinations in comparison to the conventional chaotic routing method and the shortest path routing.

1. Introduction

Recently, the Internet users continue to increase by rapid spread of convenient mobile devices. The number of flowing packets in the computer network then increases and this causes packet congestion. Therefore, an efficient packet routing method that alleviates the packet congestion is greatly desired. The shortest hop method (sp-hop method) that is commonly employed as the routing method in the communication networks [1] has serious problems; if the flowing packets in the network increases, the packet congestion occurs at the hub nodes. Thus, many packets are trapped at the hub nodes, and they spend long time to arrive at destinations. As one of the routing methods to remove the packet congestion at the hub nodes, a packet routing method using chaotic neurodynamics (chaotic method) has already been proposed [3]. By using the refractory effects which are included in chaotic neurons and can realize packet routing history, the routing method [3] avoids the packet congestion better than the sp-hop method. However, if the number of packets drastically increases, the method shows poor performance because some packets are congested at the hub nodes. We then proposed a new routing strategy for routings of packets by chaotic neurodynamics with degree information in this study. From the results of numerical simulations, the proposed routing method alleviates the packet congestion at the hub nodes effectively, achieving large number of packets are transmitted to their destinations in comparison to the conventional chaotic method and the sp-hop method. Also, by reducing the number of control parameters in the conventional chaotic method, the proposed routing strategy can apply to various topologies of the computer networks.

2. A computer network model

A network model used in this study is composed by nodes and links. Each node represents a router or a host and each link represents the connections between the nodes. It has been already reported that the Internet has scale-free property; a few nodes have many links and most of nodes have small links. Therefore, we employ the scale-free networks as topologies of the computer networks.

An example of degree distribution of the scale-free networks is shown in Fig. 1.
that is the number of packet transmissions in the single iteration. In this study, we set $C$ to 1. In addition, each node has buffer in which the packets are stored. In this study, we set the buffer size to infinity. We generated $R$ packets whose generation and destination nodes are randomly determined using uniformly distributed random numbers at each iteration. Here, $R$ is defined by $R = P_g \times N$, where $P_g$ is a packet generating probability, $N$ is the total number of nodes.

3. A packet routing method using chaotic neurodynamics

First, we describe how to realize the chaotic method [3]. We assume that the network has $N$ nodes and the $i$th node ($i = 1, ..., N$) has $N_i$ adjacent nodes. Then, the $i$th node has a chaotic neural network which has $N_i$ chaotic neurons.

![Figure 2: An example of the chaotic neural network.](image)

In the chaotic method [3], the $i$th neuron expresses connection information between the $i$th node and the $j$th node. In addition, if the $i$th neuron, the $j$th neuron in the $i$th chaotic neural network, fires, a packet is transmitted from the $i$th node to the $j$th adjacent node. Figure 2 shows an example of the chaotic neural network. To realize the packet transmission determined by the firing of neuron, we first assign the mutual connection to each chaotic neuron, achieving the sp-hop connection, $\eta_j(t + 1)$, is defined as follows:

$$\eta_j(t + 1) = W - W \sum_{j=1}^{N_i} x_j(t). \quad (3)$$

In Eq. (3), $W > 0$ is the control parameter, $N_i$ is the number of adjacent nodes of the $i$th node. In Eq. (3), if many neurons fire, the value of the second term in the right hand side of Eq. (3) increases. Finally, the output of the $i$th neuron is defined as follows:

$$x_j(t + 1) = f(\xi_j(t) + \zeta_j(t) + \eta_j(t + 1)), \quad (4)$$

where $f$ represents the sigmoid function that is defined as follows:

$$f(y) = \frac{1}{1 + e^{-y}}. \quad (5)$$

In Eq. (4), if $x_j(t + 1)$ is larger than 1/2, the $j$th neuron fires, and the packet is transmitted from the $i$th node to the $j$th adjacent node. If multiple neurons fire, we set a neuron that has maximum internal state fires.

![Figure 3: Distance information from the $i$th node.](image)
We evaluate the performance of the sp-hop method and chaotic method [3]. We repeated packet transmission for $T = 2000$. The distance between the nodes is set to 1. We set the parameters in Eqs. (1), (2), (3) and (5) as follow: $\beta = 1.2$, $\alpha = 0.045$, $k = 0.98$, $\theta = 0.5$, $W = 0.05$ and $\epsilon = 0.5$.

In these simulations, we used the packet congestion rate and the average number of receiving packets as evaluation indices for routing methods. The packet congestion rate represents degree of congestion in the network. The packet congestion rate is defined as follows:

$$\gamma = \frac{C}{R} \frac{\sum_{t=1}^{T}(h(t) - h(t-1))}{T},$$

where $C$ is transmission capacity at every node, $R$ is the number of generating packets at each iteration, $h(t)$ is the number of packets in the network at the $t$th iteration.

The network is congested if the packet congestion is close to 1. The packet congestion rate and the average number of receiving packets by the sp-hop method and chaotic method are shown in Figs. 4 and 5.

![Figure 4: Packet congestion rate by the sp-hop method and the chaotic method.](image)

In Fig. 4, although the packet congestion rate by the sp-hop method starts increasing when $R$ becomes 4, that by the proposed method start increasing at $R = 7$. From this result, the chaotic method can avoid congestion in comparison to the sp-hop method.

![Figure 5: Average number of receiving packets by (a) the sp-hop method and (b) the chaotic method ($R = 15$).](image)

In Fig. 5, the chaotic method has smaller number of receiving packets at the hub nodes than the sp-hop method. However, even if the chaotic method can remove the packet congestion at hub nodes, many packets are still trapped at these hub nodes. We then expect that the performance of the chaotic method will be enhanced if the packets trapping at hub nodes are successfully avoided. We then propose the routing strategy for routing packets using chaotic neurodynamics with degree information in the next section.

### 4. A routing strategy using chaotic neurodynamics with degree information

From the previous results, we clarified that the chaotic method cannot remove the congestion at the hub nodes, if the number of packets increases. In order to improve this problem, we change the distance information as follows:

$$\xi_{ij}(t+1) = 1 - \frac{d'_{ij} + d'_{ip(t)}}{\sum_{k=1}^{N} (d'_{ij} + d'_{ip(t)})}.$$  (7)

In Eq. (7), $d'_{ij}$ is the shortest path length using degree information from the $i$th node to the $j$th adjacent node, $d'_{ip(t)}$ is the shortest path length from the $j$th adjacent node to the $g(p(t))$. In Eq. (7), distance from the $i$th node to the $j$th node is determined using the degree of the $j$th node. Conversely, the distance from the $j$th node to the $i$th node is determined by the degree of the $i$th node (Fig. 6). By using the degree information as the distance between the nodes, we expect that the packets are hardly to be transmitted to the hub nodes.

![Figure 6: Distance between the nodes using degree information.](image)

The chaotic method [3] uses a control parameter $\beta > 0$, but our proposed method remove this parameter by regarding $\beta$ as 1. We also change $d_i$ used in Eq. (1) to the sum of the path lengths among the adjacent nodes because Eq. (1) has the possibility to take a negative value if the distance from the adjacent node is larger than the network diameter. By implementing Eq. (7) only, and we set the $j$th node which has maximum value among all the adjacent nodes as the transmitting node, we can realize a sp-degree routing method. Next, we introduce the refractory effect defined by Eq. (2) to the $i$th neuron. As compared to the chaotic method [3], functionality of the refractory effect is completely different. Although the packets are easily to be transmitted to the nodes which have small degree using Eq. (7), the packet can be transmitted to the hub nodes using the refractory effect in this method. Equation (8) represents a modified the mutual connection.
\[ \eta_j(t + 1) = \frac{1}{\sum_{i=1}^{N_j} x_{ij}(t)}. \] (8)

The chaotic method [3] uses a control parameter \( W > 0 \), but our proposed method removed this parameter. Finally, we determine the output of the \( i \)th neuron in the same way used by the chaotic method [3].

We evaluate the performance of the proposed method using the sp-hop method, chaotic method and sp-degree method. Experimental conditions are the same as those used in the Section 3.

Parameters of the proposed method were set to \( \alpha = 2.2 \times 10^{-4} \), \( k = 0.99 \). The packet congestion rates for the scale-free networks by the routing methods are shown in Fig. 7.

\textbf{Figure 7: Packet congestion rates by the routing methods.}

In Fig. 7, the proposed method shows smaller congestion rate than the other routing methods. Figure 8 shows the average number of receiving packets by the sp-degree method and our proposed method.

\textbf{Figure 8: Average number of receiving packets of (a) sp-degree and (b) proposed methods \((R = 15)\).}

In Fig. 8, the sp-degree method and the proposed method decentralize the packets at the hub nodes compared with that by the sp-hop and the chaotic methods (Fig. 5). In addition, the proposed method can remove the packets at the hub nodes in comparison to the sp-degree method.

Finally, we investigate the average difference of the number of receiving packets by the sp-degree method and the proposed method. Figure 9 shows the average difference of receiving packets by the sp-degree and the proposed methods.

\textbf{Figure 9: Average difference of receiving packets by the sp-degree method and the proposed method \((R = 15)\).}

In these simulations, we changed the distributions for generation packets for 20 times and averaged the results. In addition, we set \( R \) to 15. In Fig. 9, the nodes in the proposed method receive more packets than the ones in the sp-degree method if the value is positive. Conversely, the nodes in the sp-degree method receive more packets than the proposed method if the value is negative. From Fig. 9, the proposed method can use hub nodes to transmit the packets in comparison to the sp-degree method. These transmissions can diversify the path selections for routing the packets realized by the chaotic neurodynamics, resulting in alleviating packet congestion.

\section{5. Summary}

In this study, the packet routing method using chaotic neurodynamics method with degree information is proposed to remove the packet congestion at hub nodes. From the numerical experiments, the proposed method removes the packet congestion by diversifying the packet transmissions using low-degree nodes. Furthermore, we remove some parameters used in the conventional chaotic routing method, achieving easy setting of optimum values for effective routings. In future works, the performance evaluation by changing packet generating distribution and the transmission capacity of the each node will be conducted the research of T.K. was partially supported by a Grant-in-Aid for Young Scientists (B) from JSPS (No.16K21327).

\section{References}


Performance Evaluation and Analysis of Chaotic CDMA Considering Synchronization Acquisition

Kohei Ito†, Thong Huynh†, Kaori Kuroda†, Mikio Hasegawa†

†Department of Electrical Engineering, Tokyo University of Science
6-3-1 Nijjuku, Katsushika-ku, Tokyo, 125-0051, Japan
Email: {kohei, huythongtc}@haselab.ee.kagu.tus.ac.jp,
kuroda, hasegawa}@ee.kagu.tus.ac.jp

Abstract—Chaotic sequences have been shown to be effective for Direct Sequence / Code Division Multiple Access (DS / CDMA). Because the spreading codes generated by the chaotic sequences have negative autocorrelation, the chaotic sequences can minimize interferences of the DS/CDMA system. However, the error of synchronization acquisition due to the negative autocorrelation characteristics in chaotic sequences is not considered. Therefore, we study the bit error rate (BER) of the Chaotic sequence in DS/CDMA system by considering the synchronization acquisition. As a result, optimal autocorrelation is changed compared to conventional results. More specifically, we derive both simulation analysis and theoretically analysis to find the optimal autocorrelation to minimize error caused by negative autocorrelation in chaotic sequence with taking synchronous acquisition.

1. Introduction

In conventional DS/CDMA, the spectrum is spread into wide band by using different random code, called spreading code. For example, M sequences, gold sequences have been used as the spreading code in conventional systems. In this paper, we consider other types of spreading code generated by chaotic dynamical systems. On contrary with the spreading code generated by the god sequence which has zero cross-correlation, the spreading code generated by the chaotic system has negative cross-correlation. It has been shown in previous studies that the chaotic codes with negative autocorrelation can help to reduce the multiple access interference of the DS/CDMA system in comparison with the gold sequences [1]-[3].

However, the chaotic code generates the negative autocorrelation between the spreading codes which affects to the synchronization acquisition. In these aforementioned works, the influence of the synchronization acquisition has not been considered.

In this paper, we evaluate performance of chaotic code by taking into account both the synchronization acquisition error caused by negative autocorrelation and the negative cross correlation due to the chaotic sequences. Then, we analyze the simulation result theoretically.

In the following, DS/CDMA by chaotic sequence and these conventional researches are explained in Section 2. Then, performance analysis considering degradation of autocorrelation characteristic in DS/CDMA is shown in Section 3. Then, theoretical analysis is shown in Section 4. Finally, summary is shown in Section 5.

2. Code Division Multiple Access by Chaotic Sequence

2.1. DS/CDMA

DS/CDMA is a communication scheme that spreads spectrum into wide band by using different random code, called spreading code, for each terminal. In other words, each transmitter divides the data sequence into certain time interval (Chip length), and multiplies each chip by spreading code. Due to that, each terminal is able to transmit data simultaneously in the same frequency band. In order to reduce multiple access interference, the low cross-correlation sequences for the spreading code in DS / CDMA are needed. For example, Gold sequences are used in conventional systems.

2.2. Chaotic CDMA

Chaotic sequences have been shown to be effective for DS/CDMA, when the synchronization of the transmission timing is not taken between each transmitter, that is the asynchronous DS/CDMA. Optimal spreading code, which minimizes interferences on the asynchronous DS/CDMA, can be generated by the chaotic maps.

In [1]-[3], the effectiveness of spreading code with negative autocorrelation has been shown in the asynchronous DS / CDMA. In [2], spreading code with a negative autocorrelation has been theoretically analyzed on asynchronous DS / CDMA by calculating the multiple access interference. The Fig.1 shows the models of two users using spreading code X and Y.

In Fig.1, l and ε are the time difference of the chips. When ε is zero, it corresponds to a chip-synchronous CDMA, and when ε is not zero, it is a chip-asynchronous CDMA.
The multiple access interference between code X and Y is expressed by following equation,

\[ I = (1 - \varepsilon) R_{N}^{E/O} (l; X, Y) + \varepsilon R_{N}^{E/O} (l + 1; X, Y) \]  \hspace{1cm} (1)

where \( R_{N}^{E/O} (l; X, Y) \) is even/odd cross-correlation, which is expressed by following equations,

\[ R_{N}^{E} (l; X, Y) = \sum_{n=0}^{N-l-1} X_n Y_{n+l} + \sum_{n=0}^{l-1} X_{n+N} Y_{n} \]  \hspace{1cm} (2)

\[ R_{N}^{O} (l; X, Y) = \sum_{n=0}^{N-l-1} X_n Y_{n+l} - \sum_{n=0}^{l-1} X_{n+N} Y_{n} \]  \hspace{1cm} (3)

The value of n-th chip of the spreading code \( X_n \), is defined as follows,

\[ X_n = \{-1, +1\} \hspace{1cm} (0 \leq n \leq N - 1) \]  \hspace{1cm} (4)

The spreading factor, and, \( n \) is n-th chip value of the spreading code \( (0 \leq n \leq N - 1) \).

Here, we assume that the spreading code of each user is generated by the binary Markov chain model as in Fig.2.

\[ \begin{align*}
1 - \frac{\lambda}{2} & \hspace{1cm} 0 \leq x(t) \leq 1 + \frac{\lambda}{2} \hspace{1cm} (1) \\
1 + \frac{\lambda}{2} & \hspace{1cm} 1 + \frac{\lambda}{2} < x(t) \leq 1 \hspace{1cm} (2) \\
\frac{1 - \lambda}{2} & \hspace{1cm} x(t) < 0 \hspace{1cm} (3)
\end{align*} \]

where \( E[X] \) denotes the expectation of X. The equations (5) and (6) are corresponding to the autocorrelation of the spreading code shifted \( l \) chips. In other word, by changing \( \lambda \), auto-correlation of the spreading code can be changed.

Using these equations, we can find the multiple access interference by using the equation (1). When N is large enough, the expectation of multiple access interference can be calculated as follows,

\[ E \left( \frac{(l \lambda)^2}{\sqrt{N}} \right) = \frac{2(1 + \lambda + \lambda^2)}{3(1 - \lambda^2)} \]  \hspace{1cm} (9)

From (9), it can be clarified that the multiple access interference is minimized, when \( \lambda = -2 + \sqrt{3} \). It has shown that the spreading codes with parameters of \( \lambda = -2 + \sqrt{3} \) is able to reduce multiple access interference in the asynchronous DS/CDMA. In other word, cross correlation between users is reduced by using spreading code with negative autocorrelation.

To generate such spreading code, we use the chaotic map, called Kalman map. The Kalman map generates the spreading code according to the following equations,

\[ x(t + 1) = \frac{2x(t)}{1 + \lambda} \hspace{1cm} (0 \leq x(t) \leq \frac{1 + \lambda}{2}) \]  \hspace{1cm} (10)

\[ x(t + 1) = \frac{2x(t)}{1 - \lambda} - \frac{2\lambda}{1 - \lambda} \hspace{1cm} (1 + \frac{\lambda}{2} \leq x(t) \leq 1) \]  \hspace{1cm} (11)

\[ x(t + 1) = \frac{2x(t)}{1 - \lambda} - \frac{2}{1 - \lambda} \hspace{1cm} (1 \leq x(t) \leq \frac{3 + \lambda}{2}) \]  \hspace{1cm} (12)

\[ x(t + 1) = \frac{2x(t)}{1 + \lambda} - \frac{2\lambda - 2}{1 + \lambda} \hspace{1cm} (\frac{3 + \lambda}{2} \leq x(t) \leq 2) \]  \hspace{1cm} (13)

When \( 0 \leq x(n) \leq 1 \), spreading code \( X_n = -1 \). And, when \( 1 \leq x(n) \leq 2 \), spreading code \( X_n = +1 \). Fig.3 shows an example of the Kalman map.
In addition, we conduct the simulation to verify the our analysis in Fig.4. The horizontal axis is \( \lambda \) a parameter to determine the autocorrelation of spreading code. The vertical axis is the expectation of the multiple access interference. Here, the red line shows the expectation of the multiple access interference by calculation. Blue dot shows the multiple access interference by simulation using chaotic codes that is made by the Kalman map on spreading factor 511. It can be seen that the blue dot has a similar tendency to the red line. Furthermore, we also confirm that the interference is minimize when \( \lambda = -2 + \sqrt{3} \).

![Fig.4: Expectation of the multiple access interference.](image)

The effectiveness of the spreading code made by chaotic map in the asynchronous DS / CDMA is also shown in the simulation [3].

3. Performance Evaluation Considering Synchronization Acquisition

3.1. Influence of Autocorrelation in DS/CDMA

In previous section, optimal \( \lambda \) that minimizes the cross correlation between each spreading code has been found. Chaotic code with negative autocorrelation reduces multiple access interference. However, this analysis does not consider the effect of synchronization acquisition caused by negative autocorrelation. Generally, in DS/CDMA, the received signal is passed through to correlator, and then, determining the peak value to capture synchronization point. Then, if the spreading codes have negative autocorrelation, errors due to the occurrence of the second peak are increased.

Fig.5 shows an example of the autocorrelation properties of Gold code and chaotic code, when \( \lambda = -2 + \sqrt{3} \), on spreading factor 127.

From Fig.5 (b), we observe that the second peaks are occurred in the point shifted by one chip from the peak. Synchronization acquisition errors increase because of these second peaks.

![Fig.5: Example of the autocorrelation properties.](image)

In the following, we evaluate performance of chaotic code by considering the degradation of synchronization acquisition caused by negative autocorrelation, and also the cross correlation between each spreading code.

3.2. Simulation of Chaotic CDMA Considering Synchronization Acquisition

3.2.1. Simulation Model

In this paper, we evaluated the bit error rate (BER) of the DS/CDMA with synchronization acquisition by computer simulation. Fig.6 shows the simulation model.

![Fig.6: Simulation model.](image)

There are one pair transmitter and receiver. The terminals are the interference sources which generate the interference to the receiver. All the terminals adjust their transmit power in order to get same receive power at the receiver. At the receiver side, the received signal is passed through to correlator, and determined the peak value that is positive or negative. Then, demodulated signal is compared with the transmission signal to compute the BER. Simulation parameters are shown in Table.1.
### Table 1: Simulation parameters of Fig.6.

<table>
<thead>
<tr>
<th>Method</th>
<th>Asynchronous DS/CDMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel Noise</td>
<td>Nothing</td>
</tr>
<tr>
<td>Spreading Code</td>
<td>Chaotic code</td>
</tr>
<tr>
<td>Code Length</td>
<td>127</td>
</tr>
<tr>
<td>Number of User</td>
<td>5, 6, 7, 8, 9, 10</td>
</tr>
</tbody>
</table>

#### 3.2.2. Simulation Results

The simulation result is shown in Fig.7. The vertical axis represents the BER, and the horizontal axis represents the value of $\lambda$. We see from Fig.7 that BER is minimize when nearby $\lambda = -0.2$ in each line. As compared with conventional research, optimal lambda approaches zero when considering synchronization acquisition.

### 4. Theoretical Analysis

This section describes why the optimal lambda changes from $\lambda = -2 + \sqrt{3}$ to nearby $\lambda = -0.2$.

The autocorrelation function of chaotic code is represented by using lambda in Fig.8.

It can be seen from this figure that second peak is occurred when $l = -1$ or 1. This second peak is the cause of error, when synchronization acquisition is considered. Then, if negative autocorrelation is high, the BER of the system is increasing. However, from equation (9), it is shown that the effectiveness of the negative autocorrelation in order to reduce multi access interference by lowering cross correlation. Therefore, it is necessary to minimize the influence of both autocorrelation and cross correlation, as shown in the following objective function (14),

$$
\min \left\{ a|\lambda| + \beta \frac{2(1 + \lambda + \lambda^2)}{3(1 - \lambda^2)} \right\},
$$

where, $\alpha$ and $\beta$ is a parameter that determine the weight of each term. If the networks have less interference, the value of $\alpha$ becomes larger to reduce the BER caused by the synchronization acquisition. Otherwise, if the networks have high interference, value of $\beta$ becomes larger to against the multiple access inference.

It can be seen from objective function (14) that the optimal lambda approaches zero when considering synchronization acquisition. This is consistent with the simulation in Fig.7.

### 5. Conclusion

In this paper, we study the DS/CDMA where the spreading codes has been generated by the chaotic sequence. We focus on the influence of negative autocorrelation of the spreading code which has not been considered in previous works. More specifically, we derive both simulation analysis and theoretically analysis to find the optimal autocorrelation to minimize error with taking the synchronous acquisition caused by negative autocorrelation in chaotic sequence.

As future prospects, it is necessary to clarify the cause-and-effect relationship of the number of interference and parameters $\alpha$ and $\beta$.

This work was partially supported by JSPS KAKENHI Grant Number 26330117.

### References

High efficient THz-TDS system using laser chaos and super focusing with metal V-grooved waveguides

Fumiyoshi Kuwashima†, Takuya Shirao†, Toshihiro Kishibata†, Yusuke Akamine†, Kazuyuki Iwao†, Manatu Ooi†, Masahiko Tani†, Kazuyoshi Kurihara†, Kohji Yamamoto†, Takeshi Nagashima‡, Makoto Nakajima‡, and Masanori Hangyo§

†Department of Electrical and Electronic Engineering, Fukui University of Technology, 3-6-1Gakuen, Fukui-shi, Fukui 910-8505, Japan
‡Research Center for Development of Far-Infrared Region, University of Fukui, 3-9-1 Bunkyou, Fukui-shi, Fukui 910-8507, Japan
§Faculty of Education and Regional Studies, University of Fukui, 3-9-1 Bunkyou, Fukui-shi, Fukui 910-8507, Japan

Abstract—Generation of a wide-range THz waves are investigated from a photoconductive antenna excited using a chaotic oscillation multimode semiconductor laser with optical delayed feedback by an external mirror. The properties of the generated THz wave are compared with those of a case excited by a CW steady state laser. The stable THz wave is obtained from the multimode-laser diode excited photoconductive antenna by using a laser chaos. For a high sensitive detection, a metal V-grooved waveguide (MVG) is also used. The 1.6 times signal is detected using MVG compared with conventional system using Si lens.

1. Introduction

Generation of a stable wide-range THz Wave using a chaotic oscillation in a multimode semiconductor laser with an optical delayed feedback by the external mirror is investigated. A mode-locked Ti:sapphire laser is frequently used to excite the Voltage-biased photoconductive antenna (PA). But it is a high cost system. A multimode semiconductor laser is also used to excite the antenna1-2). This system is cheap but a spectrum of generated THz wave is essentially limited spectrum with a frequency interval between longitudinal modes of a semiconductor laser. And also time series of THz wave is not stable since mode hops in multimode semiconductor lasers suddenly occur.

We propose to use a chaotic oscillation of a semiconductor laser in order to obtain stable cheap continuously wide range THz wave. And a Metal V-grooved Wave guide (MVG) is also used to detect the THz waves effectively in this paper.(Fig.1)
Experimental setup is shown in Fig. 2. A semiconductor laser (780nm, ROHM, RLD78PPY6) is operated longitudinally multimode with a frequency interval of 43GHz between longitudinal modes without an external mirror (M3) under the condition of Iop (operation current) ≤ 120mA. The output power is fed back into laser via the external mirror(R3). Fed back rate is denoted by the effective reflectivity R3(eff). The emitter bowtie PA was applied with an AC voltage of 100 Vpp with a frequency of 40 kHz for lock-in detection. The sub-THz radiation which traveled in free space was focused on the detector PA. The photocurrent induced in the detector bowtie PA was detected by the lock-in amplifier with a time constant of 300 ms. The signal is obtained as a function of the delay time is a cross correlation between the sub-THz wave electric field and the exciting laser intensity.

3. Experimental Results

\[ T_{op}=25.0 \, ^\circ C, I_{op}=130 \, \text{mA}, \Delta \nu = 42.6 \, \text{GHz} \]

![Signal vs. Time for Si Lens, MVG, and MVG with R3(eff) = 0%](image)

**Fig. 3 THz time series (a) Si lens, R3(eff) = 5\%, (b) R3(eff) = 5\%, (c) MVG, R3(eff) = 0\%.**

The time series of generated THz wave are shown in Fig. 3. R3(eff) is fixed 5\% in Fig. 3(a) and (b). From top to bottom, THz waves are detected by using (a) Si lens, (b) MVG, and (C) MVG, R3(eff) is 0\%. Since the LT-grown substrate is opaque to 780nm laser light, the THz waves are irradiated from the back side of PA. However the THz waves are diffused in the substrate of PA, the 1.6 times signal is detected using MVG compared with conventional system using Si lens.

4. Conclusion

We demonstrated that the generated THz signals are stabilized and enhanced 10 times using laser chaos. And using MVG, the detected THz signals are enhanced 1.6 times owing the super focusing in the THz region.

References

Abnormal Cardiac Conduction in Acquired Heart Disease: a Simulation Study

Kunichika Tsumoto† and Yoshihisa Kurachi‡

†Division of Molecular and Cellular Pharmacology, Graduate School of Medicine, Osaka University
2–2 Yamada-oka, Suita, Osaka 565-0871, Japan
Email: tsumoto@pharma2.med.osaka-u.ac.jp, ykurachi@pharma2.med.osaka-u.ac.jp

Abstract—The voltage-gated sodium (Na) channels, playing key roles in the action potential (AP) initiation and the propagation, alters the distribution within a myocyte in congenital and acquired disorder. Here, we show that the Na channel distribution at the subcellular level may be partly contributed to the conduction disorder and the initiation of reentrant tachyarrhythmia by using computer simulation. These results suggest that the function of ion channels as well as the subcellular distribution may be partly responsible for the occurrence of lethal arrhythmia.

1. Introduction

The Cardiac Arrhythmia Suppression Trial (CAST) [1, 2] showed that the risk of arrhythmia-related death was increased in patients with old myocardial infarction, although Na channel blockers like class I antiarrhythmic drugs reduced premature ventricular contractions (PVCs), which degenerate into tachyarrhythmias. However, it remains controversial whether the poor prognosis is due to the negative inotropic and/or proarrhythmic effects of Na channel blockers.

It has been theoretically confirmed that a hypothetical conduction mechanism based on the microstructure of ventricular myocytes, i.e., electric field (EF) mechanism; the electrical communication between the adjacent cells was mediated by the change in the large negative extracellular potential induced at the narrow intercellular cleft space facing the intercalated discs (IDs), i.e., junctional membranes (JMs), serves as one of the homeostatic mechanisms of excitation conduction under conditions of reduced gap junctional coupling in the diseased heart [3, 4]. Here, we extend this idea to the proarrhythmic mechanism under the conditions of 5-day infarcted canine heart of which gap junctional coupling is not reduced yet [5]. Baba et al. [6] found that the subcellular redistribution of Na channels occurs at the ischemic border zone (IBZ) of 5-day infarcted canine ventricles, leading to marked decrease of Na channel expression in the lateral membrane (LM) of cardiomyocyte. The aim of this study was to reveal the role of the subcellular Na channel redistribution in proarrhythmic effects of Na channel blockers on ventricular arrhythmias under ischemia. The combined effects of the subcellular redistribution of Na channels and the Na channel blockade on excitation conduction were investigated by altering the subcellular Na channel distribution of each myocyte in a myofiber model, and proarrhythmic effects of class I antiarrhythmic drugs on infarcted ventricular tissue were evaluated on the basis of the conduction properties in the simulated myofibers.

2. Methods

2.1. Myofiber model

We constructed a myofiber model comprising of 100 ventricular myocytes, each of which was 100 µm in length and 22 µm in diameter (Fig.1A). The myocytes were electrically connected with both gap junctions and the EF...
mechanism [3, 4] (Figs.1B). On the basis of previous experimental data [5] reporting that the intercellular gap junctional conductance ($G_j$) of the IBZ in 5-day infarcted canine heart did not differ significantly from that of the non-ischemic zone (NZ), we employed the same $G_j$ in both the IBZ and NZ myofiber models: 2.534 $\mu$S. The radial cleft conductance ($G_r$) [3] and series axial cleft conductance ($G_a$) were 0.25 $\mu$S and 33.8 mS, respectively [4].

Each myocyte in the myofiber model comprised 3 segments: 2 for the JM (i.e., post- and pre-JMs), and the other one for the LM (Fig.1C). The membrane segments were represented by a modified O’Hara–Rudy dynamic (ORd) model [7, 8].

### 2.2. Myocardial ring model

To evaluate effects of Na channel blockade on the vulnerability of a myocardial ring model comprising an NZ and IBZ (Fig.1D, left) to PVCs, additional simulations were performed using the S1-S2 stimulation protocol: 8 S1 stimuli with a basic cycle length of 1,000 ms were applied transmembranously followed by an S2 stimulus with various coupling intervals (Fig.1D, right). There were 100 (cells 11–110) and 200 (cells 111–110) cells in the IBZ and NZ, respectively.

### 2.3. Subcellular Na channel distribution

The Na channel conductance was 14.868 nS/pF [8]. Thus, the entire Na channel conductance of each myocyte corresponded to 1.14 $\mu$S ($G_{Na}$), which was defined as the control value. We altered the subcellular distribution of Na channels by allocating the entire Na channel conductance to each membrane segment. The Na channel conductances of the JM and LM were expressed as percentages of the $G_{Na}$- $%g_{Na,JM}$ and $%g_{Na,LM}$, respectively. Thus, the entire Na channel conductance ($%g_{Na,JM+LM}$) was equal to the sum of $%g_{Na,JM}$ and $%g_{Na,LM}$.

As the Na channel current amplitude at the LM becomes about the same as that of IDs [9], we determined the subcellular Na channel distribution in the NZ. In particular, the Na channel conductance in the JM and LM of a myocyte located in the NZ were estimated to be half of the entire Na channel conductance [9].

### 2.4. Na channel blockade

Na channel blockade by the administration of class I antiarrhythmic drugs was achieved by reducing the entire Na channel conductance while maintaining the ratio of $%g_{Na,JM}$ to $%g_{Na,LM}$. The ratio of Na channel blockade was expressed as a percentage of the $G_{Na}$ ($%G_{Na}$ block).

### 2.5. Computations

Numerical calculations were performed as described previously [4]. Pacing stimuli of twice the diastolic threshold were applied to the LM segment of a myocyte located at one end of the myofiber.

### 3. Results

#### 3.1. Na channel blockade and conduction velocity

Figure 2A shows the relative ratios of conduction velocity (CV) normalized by CV in the myocardial fiber with 50$%g_{Na,JM}$ and 50$%g_{Na,LM}$, i.e., the subcellular Na channel distribution in the NZ, as a function of $%g_{Na,JM+LM}$ in each myocyte of the myofiber model. Thus, the CV under NZ myofiber condition were 52.6 cm/s. In the case of $%g_{Na,JM}$ change (Fig.2A, open squares with solid line), the %CV decreased not more than 10%.

Figure 2: Effects of subcellular Na channel distribution on CV and destabilization of AP propagation by Na channel blockade. (A), Relative ratios of conduction velocity (%CV) as a function of $%g_{Na,JM}$ or $%g_{Na,LM}$ in each myocyte of the myofiber model. (B), CV as a function of $%G_{Na}$ block. The relationship between the AP propagation patterns and $%G_{Na}$ block is represented by the bottom bars. SC, stable conduction; UC, unstable conduction (e.g., AP alternans); CB, complete conduction block. (C)–(E), Examples of AP propagation observed in NZ and IBZ myofibers.
as the \( g_{\text{Na,JM}} \) was reduced under the condition with \( 50\% g_{\text{Na,LM}} \). Meanwhile, in the case of \( g_{\text{Na,LM}} \) change (Fig.2A, filled circles with solid line), the CV decreased markedly as a function of \( g_{\text{Na,LM}} \) decrease under the condition with \( 50\% g_{\text{Na,JM}} \). On the basis of the experimentally measured CV [10] with reference to the previous immunostaining data [6], we determined the subcellular Na channel distribution in the IBZ as \( 50\% g_{\text{Na,JM}} \) and \( 4\% g_{\text{Na,LM}} \); the CV in the IBZ myofiber model was 32.9 cm/s.

Figure 2B shows CV as a function of \( G_{\text{Na}} \) block. In both NZ and IBZ myofibers, CVs decreased gradually with \( G_{\text{Na}} \) block. Figure 2C shows typical examples of the AP propagation observed in NZ and IBZ myofibers with \( 60\% G_{\text{Na}} \) block. The AP of each myocyte in the NZ myofiber with \( 60\% G_{\text{Na}} \) block was able to propagate through the myofiber (Fig.2C(i)), whereas the same \( 60\% G_{\text{Na}} \) block in the IBZ myofiber caused AP alternans (Fig.2C(ii)). The critical values of the \( G_{\text{Na}} \) block maintaining stable conduction were 88% and 54% in the NZ and IBZ myofibers, respectively (Fig.2B, bottom, red and blue bars denoted by SC, respectively). \( G_{\text{Na}} \) block increases exceeding the critical value in both NZ and IBZ myofibers caused unstable conduction. Typical examples of unstable conduction observed in the IBZ myofiber are shown in Fig.2D. The further increase in \( G_{\text{Na}} \) block caused complete conduction block (Fig.2E). In the NZ myofiber, the AP did not propagate when the entire Na channel conductance within the myocyte was reduced by 94%. Complete conduction block in the IBZ myofiber occurred with \( 88\% G_{\text{Na}} \) block.

3.2. Reentry induction in a myocardial ring model

A phase diagram of the excitation conduction in response to S1–S2 interval for a given degree of Na channel blockade is shown in Fig.3A. Open circles labeled \( a-d \) in Fig.3A correspond to Figs.3B(a)–(d), respectively. The blue region in Fig.3A represents the failure of AP induction due to the refractory period at the S2 stimulus site (Fig.3B(a)). Meanwhile, the gray region in Fig.3A denotes the bidirectional conduction from the S2 stimulus site, resulting in the collision of excitation waves (Fig.3B(b), asterisk). The red region in Fig.3A represents the conditions of reentry (i.e., counterclockwise rotation) induction shown in Fig.3B(c). The green region in Fig.3A (i.e., further increase in the \( G_{\text{Na}} \) block) denotes the conduction block occurring at both the proximal and distal borders of the IBZ (Fig.3B(d)). In contrast, in the case of \( 80\% G_{\text{Na}} \) block in the control model without an IBZ, no the S1–S2 intervals initiating reentry was observed.

4. Discussion

4.1. Major findings

The major findings of the present study are as follows: (1) a decrease in Na channels from the LM of each ventricular myocyte was a major cause of the conduction slowing in the IBZ; (2) an IBZ with the subcellular redistribution of Na channels was highly vulnerable to reentry under Na channel blockade. These findings suggest that the subcellular Na channel redistribution in the IBZ leads to decreased tissue excitability and that such a phenomenon is facilitated by Na channel blockers. Na channel blockade tended to cause a unidirectional conduction block toward the IBZ, resulting in the induction of reentrant tachyarrhythmia following PVC. Thus, the proarrhythmic effects of Na channel
blockers in patients with old myocardial infarction might be partly attributed to the ischemia-related subcellular redistribution of Na channels.

4.2. Proarrhythmic effects of Na channel blockade under ischemia

Cabo et al. found that the CV in the IBZ of 5-day infarcted canine ventricles is 36% slower than that in the NZ (29 vs. 45 cm/s, respectively)[10]. The experimentally observed conduction slowing in the IBZ [10] may be attributable to the subcellular Na channel redistribution via EF mechanism, because not a decrease in Na channels from the JM but the LM within each myocyte in the myofiber model resulted in a similar decrease in CV (37.5%; Fig.2A, filled circles with solid line).

Excitation conduction was more easily blocked by Na channel blockade in the IBZ than the NZ because of the marked decrease in the excitability of IBZ (Fig.2B). Accordingly, we speculate that subcellular Na channel redistribution together with Na channel blockade causes a unidirectional block at sites near the IBZ. Indeed, Na channel blockade with the coexistence of an NZ and IBZ widened the vulnerable period for PVCs (Fig.3A, red region) compared to the cases with $G_{Na}$ block. Furthermore, unidirectional block induced by a stimulus applied at a site near the IBZ initiated reentry (Fig.3B(c)).

On the other hand, as shown in Fig.2B, there existed a range of Na channel blockade causing unstable conductions (Figs.2C(ii) and D) between the ranges of stable conduction (Fig.2C(i)) and conduction block (Fig.2E). Such unstable conductions might be involved in arrhythmogenic mechanisms under ischemia [11]. Baba et al. have reported that the subcellular Na channel redistribution occurred heterogeneously at the IBZ [6]. Therefore, heterogeneous Na channel blockade in the IBZ might also cause an unstable conduction. Taken together, the present results suggest that even if the number of PVCs is reduced as a result of the continuous administration of class I antiarrhythmic drugs, the subcellular Na channel redistribution under Na channel blockade increases the ventricular vulnerability to PVCs, leading to the initiation of reentrant tachyarrhythmias and consequently more arrhythmic events [1, 2].

Acknowledgments

This work was supported by grants from JSPS KAKENHI Grant Numbers 24790214 and The Takeda Science Foundation to KT, and MEXT KAKENHI Grant Numbers 22136002 to YK.

References


Simulation of cardiac excitation propagation and the circulatory dynamics

Akira AMANO†, Hiromasa UTAKI†, Kosuke TANIGUCHI‡ and Yukiko HIMENO†

†Department of Bioinformatics, Ritsumeikan University
1–1–1 Nojihigashi, Kusatsu, Shiga 525-8577, Japan
Email: a-amano@fc.ritsumei.ac.jp

Abstract—In the clinical application, indices of whole body condition is required, on the other hand, detailed genome information often appears in the cutting edge of the basic medical research. To connect these information, we constructed a whole body hemodynamics model which consists of cellular contraction model, whole body blood vessel model and left ventricle model which can evaluate the effect of excitation propagation of left ventricle to the hemodynamics.

1. Introduction

Circulation system consists of heart and blood vessels. Blood vessels consist of artery and vein, and heart has four chambers: left and right atriums and ventricles. Human aortic blood pressure is carefully controlled to be in the physiological range of maximum 120 to 130 [mmHg] and minimum 80 to 85 [mmHg]. Heart pump function is based on the electrical excitation of cardiac cells which starts at pacemaker cells in the right atrium, and after excitation of left and right atrium, left and right ventricles are stimulated. These electrical activities can be observed with electrocardiogram (ECG), which shows 60 - 100 [ms] duration of QRS complex that corresponds to the whole heart activation time.

Since the blood pressure is one of the most important factor which determines the extracellular condition, there are several controlling mechanisms which maintains blood pressure to the physiological range. These control systems have different temporal scales. First mechanism is included in the cardiac cell, which is called force length relationship of the ventricular cell. The ventricular cell is known to have characteristics that the contraction force increases according to the increase of the initial cell length. By this mechanism, if the venous return volume increases, then the blood will be pumped out by the larger pressure. Second mechanism is called baroreflex which is controlled by autonomic nervous system (ANS). There are two blood pressure sensors in the mammalian body, and the sensed pressure information is sent to the ANS. The ANS then sends signal to sympathetic or parasympathetic nervous system and the corresponding neurotransmitter is released to the cardiac cells, which controls heart rate and contraction force. Third mechanism is called renin-angiotensin system which controls total blood volume and vasoconstriction.

These mechanisms are discovered by experiments based on single cell and whole body, i.e. microscopic and macroscopic experimental targets, however, experiments on mesoscopic experimental targets are still not fully discovered since the experiments are often difficult to conduct.

To evaluate the effect of known intercellular interaction on the blood pressure, we used a circulation model incorporating cardiac tissue model.

2. Hemodynamics Model

In this research, we used a human hemodynamic model consists of blood circulation model, an LV geometric model and a cardiac tissue model. The cardiac tissue model was constructed by connecting 10 cardiac cellular contraction models in the fiber direction to incorporate the effect of activation time.

2.1. Circulation Model

A circulation model proposed by Heldt et al. [1] (Heldt model) was used as a whole body circulation model. In our model, the LV compartment was replaced by the LV geometric model. To separate the effects of AT from other factors, the baroreflex model included in the Heldt model was removed. Cardiac cycle length was fixed at 1000 [ms]. In our combined model, we defined preload factor \( K_{rp} \) and used this factor to modify venous resistances.

2.2. Cardiac Cell and Tissue Model

We used the cardiac cellular contraction model proposed by Negroni and Lascano (NL08) [2] which offers good reproducibility in isometric and isotonic contraction, and also at various transient length changes, based on the good reproducibility of velocity-dependent stress decrease characteristics.

Cellular contraction stress \( (F_b \text{ [mN/mm}^2]) \) is calculated from the state transition model of the troponin system and mechanical model of the half sarcomere.

Since characteristics of the end-diastolic pressure volume relationship (EDPVR) are similar in rats [3] and humans [4], by linearly scaling the stress axis with the identical half sarcomere length axis, we used the following mammalian exponential function as a human passive elastic stress \( (F_p \text{ [mN/mm}^2]) \) model showing good agreement.
Table 1: Determined scale parameter values in the normal human hemodynamic model

<table>
<thead>
<tr>
<th>$L_0$</th>
<th>$D$</th>
<th>$K_{PL}$</th>
<th>$K_{PE}$</th>
<th>$K_S$</th>
<th>$F_b(ED)$</th>
<th>$K_{f_b}$</th>
<th>$K_{rp}$</th>
<th>$K_{su}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[μm]</td>
<td>[μm/m/s²]</td>
<td>[μN/m²]</td>
<td>[μN/m²]</td>
<td></td>
<td>[μN/m²]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.043</td>
<td>28.1</td>
<td>0.269</td>
<td>7.56</td>
<td>7.24</td>
<td>5.33·10⁻²</td>
<td>22.6</td>
<td>1.67</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Table 2: Parameters in Eq. (3) and (4).

<table>
<thead>
<tr>
<th>$Q_m$</th>
<th>$t_1$</th>
<th>$Q_{pump,rest}$</th>
<th>$K_p$</th>
<th>$K_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[μM/ms]</td>
<td>[ms]</td>
<td>[μM/ms]</td>
<td>[μM/m]</td>
<td>[μM]</td>
</tr>
<tr>
<td>3.2</td>
<td>8</td>
<td>0.03</td>
<td>0.15</td>
<td>0.2</td>
</tr>
</tbody>
</table>

with the experimental data [5, 6]. The format of this equation was based on the equation used by Shim et al. [7] and Landesberg et al. [8].

\[
F_p = \begin{cases} 
-K_{PL} \cdot \left(1 - \frac{L}{L_0}\right) & L < L_0 \\
K_{PE} \cdot \left(e^{\frac{D}{K_{su}}} - 1\right) & \text{otherwise} \end{cases} \tag{1}
\]

Note that $L_0$ [μm] is resting half sarcomere length, $D$, $K_{PL}$ [μN/m²] and $K_{PE}$ [μN/m²] are the scale parameters for the heart wall and cardiac cell passive elasticity. Parameter values were manually obtained to reproduce physiological human hemodynamics (Table 1).

Since $F_p$ is usually measured using a piece of tissue, we can consider that the characteristics of $F_p$ are compatible with the macroscopic properties. On the other hand, since $F_b$ is usually measured with a single cell or small piece of ventricular fiber in which the effective cross-sectional area is difficult to measure, measured stress may contain large scale errors. We thus introduced a scale factor, $K_s$, which is multiplied only to $F_b$ to adjust cellular contraction stress. Finally, total muscle stress $F$ [μN/m²] as used in Eq. (6) as wall stress is calculated as follows.

\[
F = K_s \cdot F_b + F_p \tag{2}
\]

Stimulation time in the NL08 model is controlled by the $Ca^{2+}$ release equation. The release and uptake of $Ca^{2+}$ by the sarcoplasmic reticulum ($Q_{rel}$ [μM/ms] and $Q_{pump}$ [μM/ms]) in the NL08 model are expressed using the following equations.

\[
Q_{rel} = Q_m \cdot (t/t_1)^4 \cdot e^{4(1-t/t_1)} + Q_{pump,rest} \tag{3}
\]

\[
Q_{pump} = K_p/(1 + (K_m/[Ca^{2+}])^2) \tag{4}
\]

where $t$ [ms] is the time parameter, $[Ca^{2+}]$ [μM] is the concentration of $Ca^{2+}$, $Q_m$ [μM/ms] is the maximum level of $Ca^{2+}$ release, $t_1$ [ms] is the interval to maximum $Q_{rel}$, $Q_{pump,rest}$ [μM/ms] is a parameter to determine $Ca^{2+}$ at rest, $K_p$ [μM/ms] is maximum value of $Q_{pump}$ and $K_m$ [μM] is the value of $[Ca^{2+}]$ for $Q_{pump} = K_p/2$. Parameter values used in Eq. (3) and (4) are shown in Table 2.

Since the heart can be decomposed into several long fiber bundles, we assumed that one fiber bundle surrounding the LV can be considered as a LV wall tissue model with an AT. We thus used a cardiac tissue model constructed by connecting 10 cardiac cellular contraction models in the fiber direction (10-cell model), which represents a fiber bundle surrounding the LV.

In this model, half sarcomere length was calculated by the average of 10 cells, and total muscle stress was assumed to be the same with 10 cells.

ATs can be altered by modifying the time parameter $t$ in the equation $Q_{rel}$ (Eq. (3)) by a constant time ($\delta_{delay}$ [ms]).

Since there are 10 cells in the tissue model, the relationship between AT and $\delta_{delay}$ becomes the following.

\[
AT = \delta_{delay} \times 9 \tag{5}
\]

As mentioned in Section 1, AT is closely related to QRS duration. QRS duration is under 100 [ms] in healthy human adults. On the other hand, under pathological conditions, QRS duration can be longer than 200 [ms] [9]. Here we assumed that LV AT is almost the same as QRS duration and defined two AT conditions by fixing $\delta_{delay}$ at 11 [ms] and 23 [ms].

2.3. Left Ventricle Model

In the hemodynamic model, the following geometric model based on Laplace’s law was used to relate LV pressure ($p_{lv}(t)$ [mmHg]), radius ($R_{lv}(t)$ [cm]), wall thickness ($h_{lv}(t)$ [mm]) and wall stress ($F(t)$ [μN/m²²]) [10].

\[
\frac{p_{lv}(t)}{h_{lv}(t)} = \frac{1.5 \cdot F(t)}{R_{lv}(t)} \tag{6}
\]

Since the primary variables in the hemodynamic model are LV pressure ($p_{lv}(t)$ [mmHg]), volume ($V_{lv}$ [mL]), half sarcomere length ($L$ [μm]), maximum wall thickness ($h_{lv}(t)$ [mm]), and wall stress ($F(t)$ [μN/m²²]), we have to provide the relationship between LV radius ($R_{lv}$ [cm]) and $V_{lv}$, between $R_{lv}$ and $L$, and between LV wall thickness ($h_{lv}(t)$ [mm]) and $F$.

We used the following reported data to define mathematical equations for the relationship between $R_{lv}$ and $V_{lv}$.

Corsi et al. [11] measured the time course of human LV volume and Sutton et al. [12] measured the time course of human LV internal radius. Also, Rodriguez et al. [13] reported the time course of canine LV volume, and Sabbah et al. [14] reported the time course of canine LV diameter.

We assumed that the relationship will be represented by the following equation, and from the reported data, we obtained
Table 3: Parameters in LV geometric model.

<table>
<thead>
<tr>
<th>$V_\gamma$</th>
<th>$K_\beta$</th>
<th>$K_\alpha$</th>
<th>$C_L$</th>
<th>$L_\beta$</th>
<th>$F_b(t_{ED})$</th>
<th>$F_b(t_{ES})$</th>
<th>$h_{lv}(t_{ED})$</th>
<th>$h_{lv}(t_{ES})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>[mL]</td>
<td>[µm/cm]</td>
<td>[µm]</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35.0</td>
<td>2.33</td>
<td>3.92</td>
<td>0.163</td>
<td>0.782</td>
<td>0.465</td>
<td>8.14</td>
<td>10</td>
<td>17</td>
</tr>
</tbody>
</table>

$K_\alpha = 3.92$.  

$$R_b(t) = \left(\frac{V_b(t) - V_\gamma}{K_\beta}\right)^{1/K_\alpha} \quad (7)$$

Note that $V_b(t)$ [mL] denotes LV volume and the same as $V_\gamma$. $V_\gamma$ [mL] is the $V_\gamma$-intercept of the relationship between $V_b$ and $R_b$.

We used the following reported data to define mathematical equations for the relationship between $R_b$ and $L$. Rodríguez et al. [13] also measured the time course of canine sarcomere length. By combining this data with the measured canine time course of internal diameter as reported by Sabbah et al. [14], we obtained a linear relationship between LV diameter $R_b$ and sarcomere length $L$ as $L(t) = C_L \cdot R_b(t) + L_\beta$.

Next, we considered wall thickness. Yun et al. [15] measured the time course of LV volume, twist angle, and wall thickness, and reported that both volume and twist angle showed relationships with wall thickness. We assumed that cellular contraction stress ($F_b(t)$ [mN/mm²]) is linearly related to $h_{lv}(t)$.

Parameter values of the LV geometric model are shown in Table 3.

3. Simulation Results

To realize the physiological hemodynamics, we need to adjust parameters included in the model. We have proposed a method to determine these parameter values by using three end-diastolic points and one end-systolic point [16]. By the method, we obtained the scale parameter values shown in Table 1.

![Figure 1: Time courses of $p_a$, $p_v$, $p_h$ and PV loops at different ATs.](image)

![Figure 1: Time courses of $p_a$, $p_v$, $p_h$ and PV loops at different ATs.](image)

Table 4: Hemodynamic parameters and parameters used to represent characteristics of the cell and cardiac cycle obtained at AT= 99 [ms] (normal AT) and 207 [ms] (prolonged AT).

<table>
<thead>
<tr>
<th>$p_h(t_{ES})$ [mmHg]</th>
<th>AT=99 [ms]</th>
<th>AT=207 [ms]</th>
</tr>
</thead>
<tbody>
<tr>
<td>112.3</td>
<td>111.1</td>
<td></td>
</tr>
<tr>
<td>$p_h(t_{ED})$ [mmHg]</td>
<td>5.81</td>
<td>6.77</td>
</tr>
<tr>
<td>$V_h(t_{ES})$ [mL]</td>
<td>47.5</td>
<td>54.0</td>
</tr>
<tr>
<td>$V_h(t_{ED})$ [mL]</td>
<td>119</td>
<td>124</td>
</tr>
<tr>
<td>$L(t_{ES})$ [µm]</td>
<td>1.0323</td>
<td>1.0604</td>
</tr>
<tr>
<td>$L(t_{ED})$ [µm]</td>
<td>1.1886</td>
<td>1.1947</td>
</tr>
<tr>
<td>SV [mL]</td>
<td>71.4</td>
<td>69.9</td>
</tr>
<tr>
<td>EF [%]</td>
<td>60.0</td>
<td>56.4</td>
</tr>
<tr>
<td>peak $p_h$ [mmHg]</td>
<td>113.6</td>
<td>112.1</td>
</tr>
<tr>
<td>max dp/dt [mmHg/ms]</td>
<td>6.18</td>
<td>3.48</td>
</tr>
<tr>
<td>$E_{max}$ [mmHg/mL]</td>
<td>3.58</td>
<td>3.03</td>
</tr>
</tbody>
</table>

Next, we performed simulation with two different ATs and observed the effects of AT prolongation on LV. Numerical values of hemodynamic parameters and parameters used to represent characteristics of the cell and cardiac cycle obtained from this simulation study are listed in Table 4. Time courses of $p_h$s for the two different ATs were superimposed on corresponding traces of aortic pressure ($p_a$) and venous pressure ($p_v$) in Fig. 1(a). We found that $p_h(t_{ES})$ and $p_h(t_{ED})$ changed by only −1.2 [mmHg] and +0.96 [mmHg], respectively. Both $V_h(t_{ES})$
and $V_{lv}(t_{ED})$ were confirmed to increase when AT was prolonged. Since both $V_{lv}(t_{ES})$ and $V_{lv}(t_{ED})$ increased by similar amounts (6.5 [mL] and 5 [mL]), SV did not change markedly. The peak value of $p_{lv}$ was unchanged, whereas max dp/dt showed a large decrease after AT prolongation. Simulation results showed that the isovolumic contraction phase is markedly prolonged and onset time of ejection phases is delayed by AT prolongation, which is supposed to be the outcome of decreased dp/dt.

From the time courses of $p_{lv}$ and $V_{lv}$, we obtained the PV loops as shown in Fig. 1(b) that have similar properties to the reported human PV loop. Decrease in max dp/dt was one of the most prominent effects of AT prolongation among all hemodynamic parameters. Decreases in values by changing AT from normal 99 [ms] to prolonged 207 [ms] were small for EF (-6%) and almost negligible (-2%) for SV. The decrease in EF was induced by the larger $V_{lv}(t_{ED})$ at longer ATs.

4. Conclusion

Since the elements of the hemodynamics model are based on the existing models, there are no additional mechanisms to the previously studied hemodynamics model. However, by using the serially connected cardiac cell model as the cardiac tissue model, a new mechanism of maintaining the hemodynamics against the excitation delay was found. Since this finding is based on the simulation model, further investigation with animal experiments will be necessary.

References


Mechanisms of sinoatrial node pacemaking: novel insights into roles of the pacemaker current $I_f$ from bifurcation analysis of mathematical models

Yasutaka Kurata†, Ichiro Hisatome‡, and Toshishige Shibamoto†

†Department of Physiology, Kanazawa Medical University, 1-1 Daigaku, Uchinada-machi, Kahoku-gun, Ishikawa 920-0293, Japan
‡Division of Regenerative Medicine and Therapeutics, Tottori University Graduate School of Medical Science, 86 Nishi-cho, Yonago 683-8504, Japan

Email: yasu@kanazawa-med.ac.jp, hisatome@med.tottori-u.ac.jp, shibamo@kanazawa-med.ac.jp

Abstract

To elucidate dynamical mechanisms of sinoatrial node (SAN) pacemaking with special focus on roles of hyperpolarization-activated current ($I_f$), we investigated influences of $I_f$ on parameter-dependent stability and bifurcations of SAN cells, whether blocking $I_f$ abolishes SAN pacemaking, and effects of $I_f$-dependent changes in intracellular Na$^+$ concentration ($Na$). Bifurcation analyses were performed for mathematical models of rabbit SAN cells. We conclude that 1) blocking $I_f$ abolishes SAN pacemaking only when cells are hyperpolarized; 2) $I_f$ does not enhance but attenuates robustness of SAN cells; and 3) enhancing effect of $I_f$ on SAN robustness is reversed by elevations in $Na$.

1. Introduction

Mechanism of sinoatrial node (SAN) pacemaking is one of the most important subjects to be elucidated in cardiac electrophysiology. Ionic mechanisms of SAN pacemaking have thoroughly been studied experimentally and theoretically. By the theoretical approach using mathematical models for rabbit SAN cells, we provided significant insights into dynamical mechanisms of SAN pacemaking and roles of sarcolemmal ionic currents such as the L-type Ca$^{2+}$ channel current ($I_{CaL}$), delayed-rectifier K$^+$ channel current ($I_K$) and Na$^+$ channel current ($I_{Na}$) [1,2].

Hyperpolarization-activated cation current ($I_f$) contributes to prevention of excess hyperpolarization [3], autonomic regulations of spontaneous activity [4], stabilization of pacemaker frequency [5], and diastolic depolarization in the periphery of intact SAN [6]. However, $I_f$ blockers did not abolish spontaneous activity of real SAN cells [6], suggesting that $I_f$ is not indispensable for spontaneous firings under normal conditions. Thus, the roles of $I_f$ in SAN pacemaking remained to be determined by the theoretical approach.

The aim of our study was to provide more profound insights into the roles of $I_f$ in SAN pacemaking in terms of nonlinear dynamics and bifurcation theory. Initiation and cessation of pacemaker activity are considered as bifurcation phenomena; bifurcation analysis provides an efficient way of understanding how individual currents contribute to pacemaker activities [1,2]. In this study, therefore, we performed bifurcation analyses for mathematical models of the central and peripheral SAN cells [2,8,9]. Bifurcation diagrams were constructed by calculating equilibrium points (EPs), limit cycles (LCs), their stability, and bifurcation points as functions of model parameters. We focused on the effects of $I_f$ on the stability of EPs and robustness of pacemaker activity against hyperpolarizing loads, and thus evaluated stability and dynamics of the model cells during injections of hyperpolarizing bias currents ($I_{bas}$), applications of acetylcholine (ACh) or electrotonic modulations by the atrium. Furthermore, we explored whether and how $I_f$-dependent pacemaking, defined as the pacemaker activity to be abolished by blocking $I_f$, is possible. $I_f$ effects were tested for both the intracellular Na$^+$ concentration ($Na$)-fixed system and Na$^+$-variable system. This study provides significant insights into the contributions of $I_f$ to EP instability and robustness of SAN pacemaking as well as how Na$^+$ influences $I_f$ effects.

2. Methods

2.1. Mathematical Formulation

2.1.1. Base models for central and peripheral SAN cells

We used the Kurata et al central [8] and peripheral [2] cell models, and the Maliksev-Lakatza model [9]. These models include 14 membrane current components. The membrane current system includes $I_{CaL}$, $I_{Na}$, $I_f$, T-type Ca$^{2+}$ channel current ($I_{CaT}$), sustained inward current ($I_o$), rapidly-activating ($I_{Kr}$) and slowly-activating ($I_{Ks}$) components of $I_K$, 4-AP-sensitive currents consisting of transient and sustained components, background currents carried by Na$^+$ and K$^+$, muscarinic K$^+$ channel current ($I_{KAC}$), Na$^+$-K$^+$ pump current ($I_{NaK}$), and Na$^+$/Ca$^{2+}$ exchanger current ($I_{NCX}$).

2.1.2. Incorporation of ACh effects on ionic currents

To investigate the bifurcation phenomena in the model cells during applications of ACh, we incorporated the formulas of Zhang et al [10] for $I_{KAC}$ and modifications of $I_{CaL}$ and $I_f$ by ACh, into the base models. $I_{KAC}$ density was assumed to be the same in central and peripheral cells.
2.1.3. Formulation of a coupled-cell model
We employed a coupled-cell model to investigate the electrotonic influences of atrial myocytes on stability and dynamics of SAN cells. A peripheral SAN cell model was connected to a passive membrane model for an atrial myocyte via the gap junction conductance (G_C) of 0–1000 nS. We used the capacitance of 134 pF and resistance of 100–900 MΩ for the atrial membrane model. A resting potential of the atrial myocyte was set equal to −80 mV.

2.2. Bifurcation Analysis
The model cells are 15- and 29-order autonomous continuous-time dynamical systems. Dynamical properties of model systems were determined by handling a set of 15 or 29 first-order, nonlinear ordinary differential equations. Numerical computations were performed with MATLAB 7.5 (The MathWorks, Natick, MA, USA).

Bifurcation parameters chosen in this study include 1) the maximum conductance of I_f (g); 2) amplitude of hyperpolarizing I_{Na}; 3) Ach concentration ([ACH]); 4) G_C and 5) I_{Na}. Detailed procedures for locating EPs and LCs, constructing one- and two-parameter bifurcation diagrams (BDs), and detecting bifurcations (determination of EP and LC stabilities) are provided previously [2]. We used 1) Newton-Raphson algorithm to locate EPs and to detect bifurcations of EPs; 2) brute-force approach using a MATLAB ODE solver, ode15s, to calculate stable LCs and arrhythmic dynamics; and 3) CL_MATCONT, a continuation toolbox for MATLAB, to locate unstable LCs and detect bifurcations of LCs. Types of LC bifurcations were determined by calculating characteristic multipliers.

3. Results
3.1. Influences of I_f on Bifurcations of SAN Cells
3.1.1. Effect on bifurcation during ACh application
We first examined the effects of I_f on bifurcation phenomena during ACh applications in the model cells. One-parameter BDs to illustrate stability and oscillation dynamics of the model cells were constructed as functions of [ACh] for different I_f values. We also constructed two-parameter BDs where g_f-dependent changes in Hopf bifurcation (HB) and saddle-node bifurcation (SNB) points were plotted; the critical ACh at which oscillation dynamics became arrhythmic via LC bifurcations were determined as functions of g_f. During [ACh] increases, EPs of the central cell were stabilized via HBs. Under the normal condition, LCs were destabilized via a period-doubling bifurcation (PDB); spontaneous firings became arrhythmic, abruptly shrunk in amplitude via another PDB, and finally vanished at the HB. In the I_f-removed cell, a LC became unstable via a Neimark-Sacker bifurcation (NSB) with emergence of arrhythmic dynamics. EPs of the peripheral cell were also stabilized via HBs during [ACh] increases. In the normal cell, LCs became unstable via NSBs, with arrhythmic dynamics emerging at these bifurcations. The critical [ACh] values to yield arrhythmic dynamics or quiescence became smaller as I_f decreased.

3.1.2. Effect on bifurcation during electrotonic modulation
We further examined the effects of I_f on bifurcations during G_C increases of the coupled-cell model. With increasing G_C increases, EPs of the peripheral cell were stabilized via HBs, with the cell coming to a rest. In both the normal and I_f-reduced cells, LCs were destabilized via NSBs with the emergence of arrhythmic dynamics.

3.2. Combined Effects of I_f and Other Currents
I_f and I_{Na} present at high density in the central and peripheral cells, respectively, may play pivotal roles in SAN pacemaking. Therefore, we examined the combined effects of I_f and I_{Na} or I_{Na} on bifurcation phenomena during hyperpolarization in the model cells. In the I_f-removed system, the critical [ACh] value to yield a stable EP was relatively low, with I_f increase not significantly enlarging the [ACh] region of unstable EP. The I_f-induced enlargements of the [ACh] and G_C region of unstable EPs were much greater in the periphery than in the center. The removal of I_{Na} shrunk the [ACh] and G_C region of unstable EPs. In the I_{Na}-removed system, the G_C region of unstable EPs was not significantly enlarged by I_f increase.

3.3. Searching for I_f-dependent Pacemaking
In the central cell under the normal condition, reducing I_f did not yield a SNB or HB at which a stable EP emerges, not abolishing LCs. Under the hyperpolarized conditions, however, blocking I_f led to 1) de novo creation of EPs at more negative potentials via SNBs, 2) destabilization of LCs via a PDB with emergence of period-2 periodic and chaotic dynamics, and 3) cessation of spontaneous activity via HBs. In the peripheral cell under the normal condition, reducing I_f did not cause EP stabilization or abolition of LCs. In the cell hyperpolarized by ACh applications, however, blocking I_f caused 1) negative shifts of steady-state potential (V_{ss}) with its stabilization via HBs, 2) destabilization of LCs via PDBs, and 3) cessation of spontaneous activity at the HB points.

3.4. Influences of I_{Na} on Bifurcation of SAN Cells
3.4.1. Enhancing I_f caused increases in I_{Na}
We further examined the g_f-dependent changes in I_{Na} at EPs and during spontaneous firings in the I_{Na}-variable system by constructing one-parameter BDs in which I_{Na} at EPs and during spontaneous oscillations were plotted as functions of g_f. The values of I_{Na} at EPs and during spontaneous firings became higher with increasing g_f.

3.4.2. Increasing I_{Na} shrunk parameter regions of unstable EPs and stable LCs
Because I_{NCX}, I_{NKA} and other Na^+ fluxes depends on I_{Na}, stability and bifurcations of SAN cells during g_f changes
may be affected by concomitant variations in Na. We therefore investigated how the parameter Na affects stability and bifurcations of EPs and LCs in the Na-fixed system by constructing two-parameter BDs for Na and \( g_{\text{Ca}} \). The \( g_{\text{Ca}} \) regions of unstable EPs and rhythmic firings dramatically shrank with increasing Na.

3.4.3. \( N_a \)-dependent effects of \( I_L \) on robustness against hyperpolarizing loads

While \( I_L \) enhances SAN cell robustness in the Na-fixed system at lower \( g_i \), \( I_L \)-dependent changes in Na may eliminate and reverse the enhancing effect of \( I_L \) in the Na-variable system. We examined the influences of \( I_L \) on stability of EPs and LCs, as well as their bifurcations, in the Na-variable and Na-fixed model cells during hyperpolarizing \( I_{\text{bas}} \) injections and ACh applications. Bifurcations during hyperpolarizing loads of the model cells were tested for broad ranges of \( g_i \) by constructing two-parameter BDs for hyperpolarizing \( I_{\text{bas}} \) and \( g_i \).

In the Na-variable system, the \( I_{\text{bas}} \) regions of unstable EPs and stable LCs shrank with increasing \( g_i \); spontaneous firings became unstable and arrhythmic via destabilization of LCs, and vanished via stabilization of EPs. In contrast, the unstable EP and stable LC regions of the Na-fixed system were enlarged by \( I_L \) at the relatively small \( g_i \); however, greater increases in \( g_i \) shrank the \( I_{\text{bas}} \) region of unstable EPs, whereas that of stable LCs was broadened with increasing \( g_i \).

4. Discussion

4.1. Roles of \( I_L \) in SAN Pacemaking

4.1.1. \( I_L \) itself does not destabilize an EP

\( I_L \) is expected to contribute to EP destabilization in SAN cells, like \( I_{\text{Ca,L}} \) and \( I_{\text{Na}} \) [1,2]. However, increasing \( I_L \) did not enlarge but rather shrank the parameter regions of unstable EPs. Thus, \( I_L \) itself does not contribute to EP instability, but facilitates EP stabilization. Excess \( I_L \) may counteract the destabilizing effect of \( I_{\text{CaL}} \) or \( I_{\text{Na}} \) on EPs.

4.1.2. \( I_L \) enhances SAN cell robustness to hyperpolarizing loads by preventing bifurcations

Lower conductance \( I_L \) enhanced the central SAN cell robustness against ACh-induced hyperpolarization by preventing emergence of a stable EP. Nevertheless, the \( I_L \) effect on the central cell was relatively small. The central cell robustness to hyperpolarizing loads was attenuated at higher \( g_i \), suggesting that \( I_L \) density should be small in the central region of SAN. In contrast, the peripheral cell showed continuous enlargements of the [ACh] and \( g_C \) regions of unstable EPs and rhythmic firings during \( I_L \) enhancement. Thus, \( I_L \) may enhance the robustness of peripheral SAN cells against hyperpolarizing loads by preventing EP stabilization and LC destabilization.

Previous studies revealed sinus dysrhythmia, recurrent sinus pause, and cessation of spontaneous activity in mice lacking HCN2 or HCN4 [11,12], possibly reflecting \( I_L \)-induced enhancement of the SAN cell robustness against hyperpolarizing loads. The repetitive sinus pause was prominent at low heart rates, e.g., under muscarinic stimulation or in the presence of \( I_L \) blockers [11,12]. These arrhythmic behaviors are very similar to those of the hyperpolarized model cells reproduced in this study.

4.1.3. Regional differences in \( I_L \) effects suggest different roles of \( I_L \) in center and periphery

\( I_L \)-induced enhancement of the SAN cell robustness was relatively small in the center, but relatively large in the periphery. The greater effect of \( I_L \) on the peripheral cell robustness to electrotonic loads and higher \( I_L \) density in the periphery are reasonable, because peripheral cells directly suffer the electrotonic load of the atrium and thus must be more robust to electrotonic modulations than central cells. These regional differences in the \( I_L \) effects may reflect different roles of \( I_L \) in the center and periphery of the SAN: \( I_L \) may contribute mainly to the robust pacemaking against electrotonic loads in the periphery, but mainly to the sympathetic regulation of pacemaker frequency in the center.

4.1.4. \( I_{\text{Ca}} \) and \( I_{\text{Na}} \) are involved in \( I_L \)-induced enhancement of SAN cell robustness

\( I_L \) was suggested to enhance the central SAN cell robustness against parasympathetic stimulation and the peripheral cell robustness against electrotonic loads of the atrium in combination with \( I_{\text{Ca}} \) and \( I_{\text{Na}} \) respectively. \( I_{\text{Na}} \)-induced destabilization of an EP at hyperpolarizing \( V_{\text{E}} \) would be involved in the \( I_L \)-induced enhancement of the peripheral cell robustness to electrotonic modulations. The combined effects of \( I_L \) and \( I_{\text{Na}} \) may be indispensable for prevention of EP stabilization and robust maintenance of SAN pacemaking against hyperpolarizing loads.

4.1.5. \( I_L \)-dependent pacemaking occurs in hyperpolarized cells

Our results suggest that \( I_L \)-dependent pacemaking is possible in hyperpolarized cells. Experimental reports suggested the \( I_L \)-dependent cardiac pacemaker: 1) Cs\(^+\), an \( I_L \) blocker, abolished spontaneous activity of rabbit SAN cells when hyperpolarizing \( I_{\text{bas}} \) was applied [13]; 2) the instantaneous background current in the pacemaker potential range was outward before \( I_L \) activation in rabbit SAN cells [14]; 3) HCN4-deficient mouse SAN cells were quiescent under low cAMP conditions [11]; and 4) \( I_L \)-based biological pacemakers could be created in the atrium and ventricle by HCN gene transfer [15]. Thus, bifurcations leading to \( I_L \)-dependent pacemaking may actually occur in the SAN under hyperpolarized or other non-physiological conditions.

4.2. Impacts of \( I_L \) on Robustness of SAN Pacemaking

4.2.1. \( I_L \) itself may attenuate robustness of SAN cells

Larger \( I_L \) did not enlarge but rather shrank the [ACh] and hyperpolarizing \( I_{\text{bas}} \) regions of stable LCs in the Na-
variable system. This result suggests that \( I_f \) does not necessarily enhance the robustness of SAN pacemaking. \( I_f \) may contribute mainly to the sympathetic regulation of pacemaker frequency in the center, while contributing to the robust pacemaking against electrotonic loads of the atrium in the periphery.

Overexpression of HCN-encoded pacemaker current was reported to silence biological pacemakers derived from guinea-pig atrial myocytes as a cautionary note for development of \( I_f \)-based biological pacemakers [16]. This observation may reflect that excess \( I_f \) expression yields EP stabilization and thus attenuation of pacemaker cell robustness, as suggested by our study. This finding is of particular importance for \( I_f \)-based biological pacemaker engineering where \( I_f \) could be overexpressed to several times the density of native currents [17].

4.2.2. \( I_f \) effects depend on concomitant changes in \( Na_i \)

The differences in the \( I_f \) effects between the \( Na_i \)-variable and \( Na_i \)-fixed systems come from \( I_f \)-dependent changes of \( Na_i \) in the \( Na_i \)-variable system; in the \( Na_i \)-fixed system, the parameter \( Na_i \) was shown to exert substantial influences on stability and bifurcations of the model cell via modulating \( I_{CaL}, I_{NaK} \) and \( I_{SCX} \). At lower \( g_i \), the decreased \( Na_i \) at EPs may contribute to enhancement of SAN cell robustness. On the other hand, the greater \( I_f \)-dependent shrinkage of the unstable regions in the \( Na_i \)-variable system is due to the \( I_f \)-dependent increase in \( Na_i \), as observed experimentally [4]. Thus, changes in \( Na_i \) strongly affect stability and bifurcations of SAN cells and thus must be taken into account in experimental and theoretical studies.

Acknowledgments

This work was supported by Ministry for Education, Science, Sports and Culture of Japan Grant-in-Aid for Scientific Research (C) 20590220 (to Y.K. and T.S.), and Kanazawa Medical University Grant for Promoted Research S2009-1 (to Y.K.).

References


Oscillations, bistabilities and bifurcations in a cardiac pacemaker cell model

Takumi Nagai† and Shinji Doi‡

†‡Department of Electrical Engineering, Kyoto University
Kyoto university-katsura, Nishikyo-ku, Kyoto 615-8530, Japan
Email: †nagai@rotary.kuee.kyoto-u.ac.jp, ‡doi@kuee.kyoto-u.ac.jp

Abstract—Heartbeats are controlled by electrical signals, which are generated by sinoatrial node cells. The temporal variation of the signals is described by nonlinear ordinary differential equations, and the Zhang model is one of the most well-known models of the cardiac pacemaker cell. However, the model exhibits bistability so that the model is not adequate for a model of pacemaker cells. In this paper, we perform bifurcation analysis of the Zhang model by varying various conductances of ion channels in order to improve the model suitably. These results suggest that ion currents which can modify the model are L-type calcium current, T-type calcium current, and background sodium current.

1. Introduction

The sinoatrial node is a cardiac pacemaker. The sinoatrial node cells periodically generate electrical signals, and conduct the signals to other cardiac tissues. These electrical signals are called action potentials, which are mainly related to ion channels in cell membranes. When the ion channels open, the specific ions pass through them. This process changes the membrane potential, which is a difference of electrical potential between the inside and outside of cell membrane. However, if the process becomes abnormal, it causes heart disease.

The action potential of sinoatrial node cells are described by Hodgkin–Huxley-type models [1, 2, 3]. The Zhang model is composed of the nonlinear ordinary differential equations with fifteen variables [2, 3]. In addition, the model enables us to take account of differences between center and periphery cells, which compose the sinoatrial node. Therefore, we can analyze the pacemaker activities of sinoatrial node cells in detail. However, the model exhibits bistabilities in the normal condition so that the model is not adequate for a model of pacemaker cells [4, 5]. This means that the model cannot simulate the action potentials accurately. For example, cardiac cells actually receive stimuli so that cardiac pacemaker cell models need to take account of them. In the case of cardiac cell models which exhibits bistability, generation or annihilation of action potentials is changed by each external stimulus. Due to this, the model cannot take account of sudden external stimuli.

In this paper, we focus on single center cell activities of the one-dimensional(1D)-capable model [3]. This cell model is commonly used to explore activities of coupled cells [6, 7]. Many kinds of parameters for the model have different values from those for isolated cell model. We investigate the bifurcation structure of the model by varying conductances of ion channels in order to modify the model.

2. Zhang model

The Zhang model is a rabbit sinoatrial node cell model described by the Hodgkin–Huxley-type equation with fifteen variables. The variation of membrane potential \( V \) (mV) is described by

\[
\frac{dV}{dt} = \frac{-1}{C} I_{\text{total}},
\]

\[
I_{\text{total}} = I_{\text{Na}} + I_{\text{CaL}} + I_{\text{CaT}} + I_{\text{Kr}} + I_{\text{Ks}} + I_{\text{IO}} + I_{\text{ion}} + I_{\text{f}}
\]

\[
+ I_{\text{bNa}} + I_{\text{bCa}} + I_{\text{bK}} + I_{\text{NaCa}} + I_{p},
\]

where \( C \) (pF) is the membrane capacitance, \( I_{\text{Na}}, I_{\text{CaL}}, I_{\text{CaT}}, I_{\text{Kr}}, I_{\text{Ks}}, I_{\text{IO}}, I_{\text{ion}}, I_{\text{bNa}}, I_{\text{bCa}}, I_{\text{bK}}, I_{\text{NaCa}}, I_{p} \) (pA) are the ionic currents. These currents are given by the following equations:

\[
I_{\text{ion}} = c_{\text{ion}} G_{\text{ion}} f(V, \chi)(V - E_{\text{ion}})
\]

where \( G_{\text{ion}} \) (\( \mu \)S) is the maximum conductance of ion channels. For the simplicity of bifurcation analyses, we have introduced the coefficient of the maximum conductance \( c_{\text{ion}} \) whose standard value is 1.0. \( \chi \) is the gating variable, which expresses opening and closing dynamics of ion channels. Temporal variations of gating variables are described by

\[
\frac{d\chi}{dt} = \alpha_{\chi}(V)(1 - \chi) - \beta_{\chi}(V)\chi
\]

where \( \alpha_{\chi}(V) \) and \( \beta_{\chi}(V) \) are the (voltage-dependent) rate constants of the transition between open and closed states of the gate. For more details, see the reference [2, 3]. We analyzed the center cell model.
3.1. Effective parameters

Stabilities of equilibrium points and periodic solutions don’t always depend on large ion currents. For example, hyperpolarization-activated current $I_{bNa}$ has an important effect upon the automaticity of cardiac pacemaker cells although the maximum value of $I_{bNa}$ is very small. Therefore, we analyze the model as for all currents even if some of them are small.

3.1.1. The $L$-type calcium current $I_{CaL}$

$L$-type calcium current $I_{CaL}$ is large current in a center cell, and plays a main role in the depolarization of action potential. Figure 1(a-1) shows the bifurcation diagram as for $c_{CaL}$. The membrane potential $V$ in the steady state is plotted for each value of $c_{CaL}$ in the diagram. The solid and broken curves show stable and unstable solutions, respectively. The thick and thin curves present periodic solutions and equilibrium points, and HB, SN, PD denote the bifurcation points of Hopf, saddle-node, Period-doubling bifurcation.

When $c_{CaL}$ is increased from 0.00, a equilibrium point is unstable at HB1 ($c_{CaL} = 1.08$), and the unstable periodic solutions are generated. Periodic solutions become stable at DC1 ($c_{CaL} = 0.74$), and become unstable at DC2 ($c_{CaL} = 10.39$). Eventually, periodic solutions disappear at HB2 ($c_{CaL} = 10.06$), so that stable periodic orbits coexist with stable equilibrium points within the range $0.74 < c_{CaL} < 1.08, 10.06 < c_{CaL} < 10.39$. Therefore, the Zhang model is suitable as a cardiac pacemaker cell model in the condition of $1.08 < c_{CaL} < 10.06$.

If membrane potentials are abnormal, the change of a model parameter is not adequate. Therefore, we evaluate the maximum systolic potential and the period of the Zhang model based on Fig. 1(a-2). The result of the simulation in normal conditions is Fig. 1(a-3), and the maximum systolic potential is 20.9 mV and the period is 336.3 ms. Furthermore, the membrane potential converges to a fixed value at $c_{CaL} = 0.70$ (Fig. 1(a-4)). In this figure, the solid line is action potential waveform in the modified condition, and the broken line is that in the normal condition. When $c_{CaL}$ is 1.08, the maximum systolic potential is 23.1 mV and the period is 338.0 ms (Fig. 1(a-5)). The maximum value of the action potential needs to be larger than 0.0 mV, so that the difference of the values is permissible. The difference of periods are also allowable.

3.1.2. The background sodium current $I_{bNa}$

Unlike $I_{CaL}$, $I_{bNa}$ is so small that it seems that $I_{bNa}$ will not affect cardiac pacemaker activities. However, as can be seen from Fig. 1(b-1), this expectation is not correct. Stable periodic orbits coexist with stable equilibrium points within the range $-0.56 < c_{bNa} < -0.53$,

![Figure 1: One-parameter bifurcation diagrams as for (a) $c_{CaL}$, (b) $c_{bNa}$, (c) $c_{CaT}$, (d) $c_{Kt}$, (e) $c_{NaCa}$, and results of the simulation at (a-3) $c_{CaL} = 1.00$, (a-4) $c_{CaL} = 0.70$, (a-5) $c_{CaL} = 1.08$, (b-2) $c_{bNa} = 0.69$, (c-2) $c_{CaT} = 0.59$.](image)
−0.36 < \( c_{\text{Na}} \) < 0.00, 0.69 < \( c_{\text{Na}} \) < 2.09. Accordingly, periodic solutions are monostable when \( c_{\text{Na}} \) decreases from 1.0.

As \( c_{\text{CaL}} \), we compare the differences between the normal condition and the modified condition. The maximum systolic potential is 20.4 mV and the period is 378.0 ms in the condition of \( c_{\text{Na}} = 0.69 \) (Fig. 1(b-2)). The difference between the values of the maximum systolic potential at \( c_{\text{Na}} = 1.00 \) and at \( c_{\text{Na}} = 0.69 \) is smaller than \( c_{\text{CaL}} \), but the period at \( c_{\text{Na}} = 0.69 \) is larger than at \( c_{\text{Na}} = 1.00 \). This means that the heart rate per minute at \( c_{\text{Na}} = 0.69 \) is 20 less than that at \( c_{\text{Na}} = 1.00 \), but this model describes a rabbit sinoatrial node cell activity so that it may not be a problem. This is the reason why \( c_{\text{Na}} \) is inferior to \( c_{\text{CaL}} \) to modify the model.

3.1.3. The T-type calcium current \( I_{\text{CaT}} \)

\( I_{\text{CaT}} \) is smaller than \( I_{\text{Na}} \), but \( c_{\text{CaT}} \) is an effective parameter to modify the model. Figure 1(c-1) shows that stable periodic orbits coexist with stable equilibrium points within the range −4.00 < \( c_{\text{CaT}} \) < −2.23, 0.58 < \( c_{\text{CaT}} \) < 4.75. Therefore, the model is suitable as a cardiac pacemaker cell model in the condition of 0.00 < \( c_{\text{CaT}} \) < 0.58 (\( c_{\text{CaT}} < 0.00 \) is physiologically impossible).

When \( c_{\text{CaT}} \) is 0.58, the maximum systolic potential is 20.1 mV and the period is 355.2 ms (Fig. 1(c-2)). The change of the value of \( c_{\text{CaT}} \) hardly affects the maximum value of action potentials. Thus, compared with \( c_{\text{CaL}} = 1.08 \), the maximum value of action potentials is closer to the value in the normal condition. On the other hand, the period at \( c_{\text{CaT}} = 0.58 \) is 10 less than in the normal condition. The influence on the period of action potentials by decreasing the value of \( c_{\text{CaT}} \) is smaller than that by decreasing the value of \( c_{\text{Na}} \). Therefore, from a viewpoint of the difference of periods, the parameter \( c_{\text{CaT}} \) is more suitable than \( c_{\text{Na}} \), but \( c_{\text{CaL}} \) is more appropriate than these.

3.2. Non-effective parameters

In this section, we introduce the parameters which are not adequate to modify the model. We show two representative examples as follows.

3.2.1. The rapid delayed rectifying potassium current \( I_{\text{Kr}} \)

The potassium currents have an effect on repolarization of the pacemaker activity, and \( I_{\text{Kr}} \) is a larger current than \( I_{\text{Ks}} \). However, the model isn’t suitable as a cardiac pacemaker cell model as \( c_{\text{Kr}} \) is any value. Figure 1(d) shows that a equilibrium point is stable with any value of \( c_{\text{Kr}} \) when \( c_{\text{Kr}} \) is increased from 0.00. On the other hand, a stable periodic solution exists between DC (\( c_{\text{Kr}} = 0.48 \)) and PD (\( c_{\text{Kr}} = 4.92 \)). Consequently, changing the value of \( c_{\text{Kr}} \) can’t improve the model.

3.2.2. The sodium–calcium exchanger current \( I_{\text{NaCa}} \)

\( I_{\text{NaCa}} \) connects with sodium ion and calcium ion movements. As can be seen from Figure 1(e), although equilibrium points become unstable and periodic solutions become stable by varying the value of \( c_{\text{NaCa}} \), this condition is physiologically impossible. Owing to this, \( c_{\text{NaCa}} \) is not suitable to improve the model.

We investigated other conductance coefficients. However, pacemaker activities are abnormal conditions as a cardiac pacemaker cell model although \( c_{Ks}, c_{Na}, c_{CaL}, c_{f}, c_{Kr}, c_{p} \) are any value.

4. Interrelation between two conductance coefficients of ionic currents

In the above section, we analyzed bistabilities of the Zhang model on each conductance coefficient of ion channels. This results indicate that \( c_{\text{CaL}}, c_{\text{Na}}, c_{\text{CaT}} \) are important to modify the model. However, the larger the value of parameter changes, the greater difference between the waveform in normal conditions and in modified conditions becomes. Thus the change of the value needs to be smaller. For this reason, we explore the adequate values of parameters to modify the model when two parameters change.

4.1. The L-type calcium current \( I_{\text{Cal}} \) and the background sodium current \( I_{\text{INa}} \)

In the previous section, increasing \( c_{\text{CaL}} \) is to make the amplitude of membrane potentials larger. Furthermore, the period of the pacemaker activities become larger because of decreasing \( c_{\text{Na}} \). Thus, we expect that the shapes of action potential in modified conditions resemble those in the normal conditions when \( c_{\text{CaL}} \) and \( c_{\text{Na}} \) are varied simultaneously. This section shows the range of parameters which makes the model adequate. Future work is needed to evaluate the shapes of membrane potentials.

Figure 2(a) shows the two-parameter bifurcation diagrams as for \( c_{\text{CaL}} \) and \( c_{\text{INa}} \). The shaded area indicate that the model is adequate as a pacemaker cell model (oscillations are stable and equilibrium points are unstable). When \( c_{\text{INa}} \) is increased from 0.00, the range between HB1 and DC1 becomes wider. Accordingly, the changes of the value of parameters are smaller to modify this model when \( c_{\text{CaL}} \) increases and \( c_{\text{INa}} \) decreases from 1.00.

4.2. The L-type calcium current \( I_{\text{Cal}} \) and the T-type calcium current \( I_{\text{CaT}} \)

As can be seen from Fig. 2(b), when \( c_{\text{CaT}} \) is increased from 1.00, the range between HB1 and DC1 becomes
wider. DC1 curve is parallel to the $c_{CaT}$-axis so that the change of $c_{CaT}$ hardly affects the contour lines of DC1 within the range $0.00 < c_{CaT} < 2.00$. On the other hand, HB1 curve runs to the upper right of the figure. Therefore, the range of $c_{Cal}$ between HB1 and DC1 becomes wider as $c_{CaT}$ is increased.

4.3. The background sodium current $I_{bNa}$ and the T-type calcium current $I_{CaT}$

Unlike $I_{Cal}$, $I_{bNa}$ and $I_{CaT}$ hardly affect the amplitude of pacemaker activities. Therefore, Fig. 2(c) is also expected to be useful to modify the model.

Fig. 2(c) shows that HB1, HB2, and DC1 curves run to the upper left of the figure. This means that the value of $c_{bNa}$ affect loci of bifurcations much the same as that of $c_{CaT}$.

5. Conclusion

In this paper, we analyzed the bifurcation structure of the Zhang model, which exhibits bistabilities in normal conditions. As a result, bifurcation structures for the single conductance coefficient of ionic currents revealed that oscillations are stable and equilibrium points are unstable by varying $c_{Cal}$, $c_{CaT}$, or $c_{bNa}$. Therefore, in this study we showed that these conductance coefficients are important for the improvement of the model.

Furthermore, we have examined two-parameter bifurcation diagram, where the bifurcation parameters are $c_{Cal}$–$c_{bNa}$, $c_{Cal}$–$c_{CaT}$ and $c_{bNa}$–$c_{CaT}$. These show that two-parameter bifurcation analysis makes the amounts of change of the bifurcation parameters smaller than the one-parameter bifurcation analyses to make the model adequate as a cardiac pacemaker cell model. These two results are key to the modification of the model.

Acknowledgement

This work was supported by the Japan Society for the Promotion of Science (24500274, 5K00337)

References


Alternans generation in the crustacean heart model: a simulation - electrophysiology study

Toru Yazawa and Hiroyuki Kitajima

Faculty of Engineering, Kagawa University
2217-20 Hayashi, Takamatsu, Kagawa, 761–0396 Japan
Email: kitaji@eng.kagawa-u.ac.jp

Abstract—Alternans-rhythm indicates an unhealthy state of the heart but this 2-beat rhythm is still required for further studies since the ionic mechanisms are not fully understood. With gathering real-world data of alternans, we characterize its ill-condition using modified detrended fluctuation analysis (mDFA) and we also argue ionic mechanisms for generating alternans based on a numerical simulation. Alternans-electrocardiograms were recorded from various “hearted” animals including humans. We found: (1) Alternans decreases the scaling exponents, which can quantify the state of the circulatory system. (2) Simulation analysis reveals that mainly K+ and/or Na+ ionic mechanisms can explain how and why the heart system generate alternans.

1. Introduction

The cardio-vascular system (CV) is a most incredible organ system in our body. It never stops to repair itself [1]. The CV is composed of a pump and a controller, i.e., the heart and the brain, respectively. The hearts are able to pump even after the isolation from the CV. Therefore, we understand that the heart has an automaticity, i.e., it has its own pace-maker. The pacing mechanism is completely intrinsic, generating electrochemical signals to keep a non-stopping blood flow throughout the body until the end of its life. However, the rhythm of the electric signals are never constant and regular, because the brain changes the rhythms according to local body’s demands, such as oxygen necessity to work out various body cell functions, locomotion and digestion for example.

When the heart beats, the intrinsic oscillator produces a signal which is a mixed/complex flow of ionic currents, flowing through the ionic channels locating in every cell membrane of the heart muscles. The pace-making is an incredible bio-chemical and bio-physical outcome generated from the complex cellular ions-proteins-water system. Due to its complexity, it is hard to identify the causes of circulatory failure, if the complex system gets a malfunction. The complex cellular ions-proteins-water system receives influences from various functional elements, which are coupled nonlinearly together, for instance, an electrical coupling between myocardial-cells, hormonal influences derived from the brain, and cardio-regulatory neuronal commands from the brain, as well as a reflexive feedback control, between the heart and the brain.

Since an electrocardiogram (EKG) was made by Willem Einthoven (the Nobel Prize in Medicine in 1924), a century have passed. But there are still so many people suffering from the cardio-vascular disease. We still need to make an effort to find out solutions to stop the life threatening disease. Our ultimate purpose throughout this study is to protect/stop the system from malfunction.

2. Alternans

One of the fatal rhythm is “the harbinger of death rhythm [2],” so called as alternans, which was documented in 1872 by a German physician Dr. Traube [3]. This 2-beat rhythm is, therefore, a well-known fatal disturbances (perturbation) in heart rhythms. However, the mechanisms for this erratic rhythm generation has not been fully understood.

Two-beat rhythm might be a sign of unhealthy state of the pumping of the heart. When does this unusual rhythm happen to us? We still do not know the reason why it is generated. In fact, Traube [3] noticed this enigmatic rhythm in human patients long ago. However, we found that it is not confined to human hearts. Here we report that alternans-like heartbeats are observable everywhere in “hearted” animals. Animals are all the same. Alternans is a fundamental phenomenon for the life on Earth. We found examples for the 2-beat EKGS recorded from various creatures, such as insects and crustaceans, as well as from human subjects. Furthermore, using mathematical simulation methods [4], we present compelling evidence for that the generation of alternans depends on the blood-concentration of potassium ions and/or the degree of membrane conductance of sodium ions. We discuss ionic bases for explaining abnormality related to interactions between potassium and sodium ionic-flow of the heart muscles.

3. Materials and Methods

3.1. Electrophysiology

Electrophysiological recording methods are described elsewhere [5]. Using our electrophysiological methods, we gathered real-world data, EKGS (Figure 1).

Heart rhythm never keeps a constant frequency. As aforementioned there is a necessity of changing the rhythms for meeting demand of the body. In healthy
individuals, the heart rate is changeable under the control of the brain. Disturbances of regulation of healthy heart rhythms are thus sometimes extremely fatal, where the heart can beat erratically.

Figure 1: EKG recordings from invertebrate animals.

3.2. Simulation

Simulation methods are described elsewhere [4]. We studied ionic currents with the numerical simulation that can see healthy/unhealthy state by changing parameter(s) in a set of numerical models.

3.3. Ethic

The heartbeats were recorded outside of a hospital, in for example university laboratories. All subjects were treated as per the ethical control regulations of Tokyo Metropolitan University.

4. Results

4.1. Crustacean isolated heart

Isolation of the heart from the body (Figure 2A) easily lead the heartbeat into arrhythmic condition (Figure 2B). Isolated hearts must die sooner or later. This is evidence that alternas is the harbinger of death rhythm, although the mechanisms remains to be elucidated.

4.2. Crustacean intact heart

Long term EKG recordings from various invertebrate animals (see Figure 1) revealed that alternans always appears before the animals die (Figure 3).

Figure 2: Regular (A) and irregular (B) heartbeats. Inset: a diagrammatic drawing of an isolated crustacean heart, simultaneous recording of pacemaker impulses (CG) and contraction force (F). Hermit crabs, *Dardanus crassimanus*.

Figure 3: EKG and heart rate of a dying crustacean, isopods, *Ligia exotica*. Two metal electrodes, 200 micrometer in diameter, were placed on the heart by penetrating them through the dorsal carapace (Figure 1). A sticky tape on a cardboard immobilized the animal. Records are shown intermitently for about one hour and 20 minutes. From H to M, the EKG and heart rates are enlarged. Small 5 arrows indicate alternans, which is observable at H-L. From Q to R, no means that the computation failed, due to a small size signal with a sporadically appearance. (After T. Yazawa et al., reference [6])

4.3. mDFA: crustacean heart alternans
Modified detrended fluctuation analysis (mDFA) was recently made [7], which can quantify either a healthy heart or a sick heart, such as ischemic failing hearts. We used mDFA method to calculate the scaling exponent (SI, scaling index) using “the heartbeat time series analysis.” mDFA can tell that a healthy heart exhibits an SI near 1.0 but a sick heart, such as dying heart exhibits a lower SI, ~0.7 for example (see detail reference [7]).

Figure 4 shows examples of the heartbeat interval time series, obtained from EKGs of a spiny lobster, one is isolated (Figure 4A) and the other one is an intact heart (Figure 4B). mDFA computation revealed that isolated hearts exhibit an SI ~0.6 and intact hearts exhibits SI ~1.0 (data not shown).

Fig. 4: Heartbeat interval time series obtained from an isolated (A) and intact hearts (B) of a spiny lobster, Panulirus japonicus. A. Alternans appeared all the way down from the first beat to the 4000th beat. B. An intact heart of a spiny lobster, Panulirus japonicus. No alternans appeared. The heart rate (shown in Hz) frequently dropped down, so-called bradycardia, which is known in normal crabs and lobsters [7]. The present mDFA revealed that the scaling exponent is normal (1.0) when the lobster is freely moving in the tank [7]. (After T. Yazawa et al., reference [6]).

Fig. 5: Alternans observable in human. A volunteer woman age 65. Upper trace, recording of finger pulses. Lower trace, heart rate. Both amplitudes alternans and intervals alternans can be seen. (After T. Yazawa et al., reference [6]).

4.4. mDFA: human heart alternans

In human, alterns also shows a low SI (Figure 5). Her SI was near 0.6. She passed away two years after this recordings. She suffered from cancer and pancreas problems. When we met her, she was physically weak and she could not walk a long distance. However, she talked with an energetic attitude. She was at first nervous because of us, but finally she got accustomed to our finger pulse testing task, and then she became relaxed. Hours later, we were surprised to note that her alternans decreased in numbers. The heart reflects the mind.

4.5. mDFA results: overview

Alternans heartbeats always decrease SI value. Figures 6, 7, and 8 show representative three “alternans hearts,” including two human hearts and one lobster isolated heart. Figure 6 shows mDFA results where all three time series exhibit a low SI while many alternans heartbeats appeared. Figures 7 and 8 show enlargements of corresponding time series in an expanded time scale (note bar periods and beat number).

4.6. Simulation

An alternans simulation study conducted by H. Kitajima has appeared elsewhere [4], where he revealed that alternans generation is caused by a shift of potassium ion concentration in blood and membrane permeability of sodium ions over the myocardium.

His conclusion was wonderfully simple, which is shown as a diagram in Figure 9. Here, one can see a parabolic relationship between potassium-ion and sodium-ion function determining the myocardial cell membrane excitability. The point A in the graph shows the state of “normal,” where both “sodium permeability” and “extracellular potassium concentration” are “normal,” i.e., a healthy range of the excitability.

Figure 6: Alternans real world data. Four example time series. A, human subject shown in Figure 5. B, lobster specimen, isolated heart shown in Figure 4A. C, human subject, male, age in his 60s.
Alternans is a sick state of the heart [2, 3]. Alternans generation was never reported before in invertebrate hearts. We found alternans in many specimens, which include crabs, lobsters, insects, and humans. Typically, all of them were found while specimens were dying (Figure 3). We conclude that the harbinger of death rhythm is observable not only in human hearts (Traube’s first documentation) but also in many “hearted” lower animals. The present study proves that, evolutionally, alternans is arising from the basic physiology mechanism, which is common to all “hearted” animals.

The mechanism is well characterized by Kitajima’s simulation [4], which explains ionic mechanism for the generation of alternans. What are the causes? It is a membrane potential shift towards a depolarizing direction (see Figure 9). An increase of blood K+ (a shift from A to B, and B to C) depolarizes the membrane according to Nernst (see Physiology text book). At B, here, a regular rhythm drastically changes to 2-beats rhythm. This correspond with a natural data (see Figure 3). From B to C, further depolarization maintains alternans state (see also Figure 3). Further depolarization returns the heart to a rhythmic state (Figure 9). But this is an extreme state, not normal. At this point, the heart enters sick and terminal state as shown in natural data (Figure 3) where the heart rate increase significantly and animals are dying.

Blood sodium concentration is high (~40 times) compared to that of potassium. Significant decrease of blood sodium ions is not realistic to happen. But if sodium permeability decreases, alternans can happen. (Figure 9, from A to D). This explains a genetically induced Na+ channel problem. The simulation worked well to understand causes of alternans, together with natural data.

Alternans is a risky state for maintaining life. mDFA can quantify this state (Figures 4, 5, and 6). Decreased SI indicates a sick state where a heart endures a condition of unhealthy (failing) heartbeats. Main causes for alternans are therefore K+ induced deadly and irreversible depolarization of heart muscles. This state is hyper K+ in blood.

Acknowledgments

The authors would like to thank NOLTA2016 organizing committee members for their fruitful suggestions and comments. This work was supported by JSPS Grants No. 23500524, and 26350508. We thank DVx Inc. Tokyo, Japan, for their support to this research since 2012 (code: 4DQ404).

References

Alternans in a Crustacean Cardiac Model

Akatsuki Abe, Hiroyuki Kitajima and Toru Yazawa

†Faculty of Engineering, Kagawa University
2217–20 Hayashi, Takamatsu, Kagawa 761-0396, Japan
Email: s16g451@stu.kagawa-u.ac.jp, {kitaji,tyazawa}@eng.kagawa-u.ac.jp

Abstract—Alternans is a kind of arrhythmia, which shows the alternation of the action potential duration (APD) or the amplitude of the action potential, and is a sign of future cardiac arrest in animal experiments. In this paper, we investigated the parameter dependence of the alternans using a mathematical crustacean cardiac model. We determined that the conductance of the calcium-dependent potassium current for the small cell is a key to generating alternans.

1. Introduction

Alternans is a beat-to-beat alternation in the action potential duration for a cardiac cell and may cause sudden cardiac death [1, 2, 3, 4]. Thus, studies of alternans using mathematical models are important to reduce the risk of sudden death. We investigated the Luo-Rudy (LR) model [5] with an external synaptic current and clarified that the alternans is generated by the period-doubling bifurcation due to increasing the value of [K]o (extracellular potassium ion concentration) [6]. [K]o affects the conductivity and the reversal potentials of several potassium currents. We were able to determine that increasing the reversal potential (E_{K1}) of the time-independent potassium current is important for the occurrence of alternans [7]. Our results showed that even though the rhythm from the pacemaker cell is normal, the alternans occurs due to the problem of the ventricular muscle.

As a next step, we have to investigate the system of making rhythmic signals, which stimulate the ventricular muscle. Here, we use a mathematical model of the crustacean heart. Because the crustacean cardiac pacemaker cells consist of a small number of neurons (four small cells (SC) and five large cells (LC), which drive muscle cells (MC)). Moreover, the network structure of the heart and the central nervous system, and the types of synapses between the SCs and LCs, the LCs and MCs, were clarified through an experiment on American lobsters [8]. Ball et al. proposed a mathematical model of the crustacean cardiac motor neurons composed of the LC and SC [9], which make signals to the cardiac muscle cells. The SCs and LCs are synchronized in each group, thus only one pair is considered. In [10], we modified some parameter values of this model to reproduce more accurate waveforms of the membrane potentials of the SC and LC. Then, we clarified that decreasing the G_{Kd} (conductance of delayed potassium current) triggers the alternans. However, the reproducibility of the experiment was limited, because the waveforms of the membrane potential for the alternans between the experiment and simulation were quite different. In particular, a long burst corresponding to the alternans for the SC was not observed in the simulation.

In this paper, we studied all combinations of the parameters (conductances of all ionic currents) and determined that a key parameter of generating a long burst is G_{KCa} (conductance of calcium-dependent potassium current). By decreasing the value of G_{KCa} for both the SC and LC, we can reproduce the firing patterns of an experiment on a hermit club.

2. Model Equation

A schematic diagram of the model proposed by Ball [9] is shown in Fig. 1. This is a two-compartment model divided by the soma and the axon. The excitatory synapses of the axon of the LC (it is not shown in Fig. 1) are connected to the cardiac muscle cells. Therefore the output in this model is the membrane potential of the axon of the LC.

![Schematic diagram of the model](https://example.com/schematic.png)

Figure 1: Schematic diagram of the model. S and L are small cell and large cell, respectively. White rectangle and circle with plus sign represent the electrical synapse or gap junction and the excitatory synapse, respectively.

We summarize the model equations described in [9]. The membrane potentials of the LC is described by:

\[
C_s \frac{dV_s}{dt} = -g_{Ls}(V_s - E_{Ls}) - g_c(V_s - V_a) - \sum I_{ls} - I_{syn} - I_{gap}
\]
\[ C_a \frac{dV_a}{dt} = - g_{La} (V_a - E_{La}) - g_c (V_a - V_s) - \sum I_i,a \]

where \( V_s \) and \( V_a \) are the membrane potential for the soma and the axon, respectively. \( C_a/C_v, S_L/a/S_L \) and \( E_{La}/E_{La} \) are the membrane capacitances, the leak conductances and the reversal potentials for the soma/axon compartment. \( g_c \) is the coupling conductance between the soma and the axon. Here, \( g_c \) with two connexons is given by

\[ g_c = (s_1 + s_2)g_{min} + s_0g_{max}, \]

where \( s_i (i = 0, 1, 2) \) is the fraction of channels in state \( S_i \):

- \( S_0 \): gate 1 open, gate 2 open,
- \( S_1 \): gate 1 open, gate 2 close,
- \( S_2 \): gate 1 close, gate 2 open.

Note that there is little evidence for the case of both gates being closed. Using the constraint \( s_0 + s_1 + s_2 = 1 \), the gate kinetics is described by

\[ \frac{ds_1}{dt} = \beta(u) - [\alpha(u) + \beta(u)]s_1 - \beta(u)s_2 \]
\[ \frac{ds_2}{dt} = \beta(-u) - [\alpha(-u) + \beta(-u)]s_2 - \beta(-u)s_1 \]

where \( u = V_i - V_a \) and

\[ \alpha(u) = \lambda \exp[-A_4(u - v_0)] \]
\[ \beta(u) = \lambda \exp[-A_0(u - v_0)]. \]

The parameter values are \( \lambda = 0.68, A_4 = A_0 = 0.10 \text{[mV]}^{-1}, g_{max} = 0.05, g_{mix} = 0.01 \text{ and } u_0 = 4.0 \text{[mV]} \) \([11]\).

The synaptic and gap currents from the SC to LC are given by \( I_{syn} \) and \( I_{gap} \), respectively. \( I_{i,S} \) and \( I_{i,a} \) are ionic currents for the soma/axon described by

\[ I_{i,S} = G_{i,S} h^p (V - E_i) \]

where \( G_i \) is the maximum conductance, \( m \) is the activation variable with exponent \( p \), \( h \) is the inactivation variable with exponent \( q \). Ionic currents are \( i = C \) transient calcium current), \( G_i \) (persistent calcium current), \( A \) (early outward potassium current), \( K_d \) (delayed outward potassium current), \( K_C \) (calcium-dependent potassium current), leak (leak current) for \( I_{i,S} \), and \( i = Na \) (transient sodium current), \( K_d \) (delayed outward potassium current), leak (leak current) for \( I_{i,a} \). The dynamics of the gating variables \( x \) (\( m \) or \( h \)) is given by

\[ \frac{dx}{dt} = \frac{x_{\infty}(V) - x}{\tau(x(V))} \]

where \( x_{\infty} \) is the steady state value and \( \tau \) is its time constant. The membrane potentials of the SC are also expressed by the similar equations, but the kind of the ionic currents is different. For detailed explanation and the values of the parameters, see [9].

**Figure 2:** Pulse trains corresponding to contraction of cardiac muscle cell for hermit club. Red and black indicate pulses of LC and SC, respectively.

### 3. Results

#### 3.1. Experiment

Figure 2 shows firing patterns of an isolated cardiac ganglion cells of a hermit club. From the observed data, we separate the firing patterns of the SC and LC. Red and black lines indicate firing timing of the LC and SC, respectively. In Fig. 2(a), bursts of the SC and LC have one to one relationship, which corresponds to a normal state. On the other hand, we observe a long burst of the SC in Fig. 2(b) and the almost two (red) to one (black) relationship appears. This state corresponds to alternans.

#### 3.2. Simulation

We studied the occurrence of a long burst of the SC for 91 \((1_{4}C_2)\) parameter planes of 14 conductances of ionic currents. We determined that a key parameter of generating a long burst is \( G_{K_{Ca}} \). Figure 3(a) shows a number of spikes in one burst when changing the value of \( G_{K_{Ca}} \) for the SC. The vertical axis indicates a maximum number of spikes in one burst after a transient period. If this number is large, we observed a long burst. From Fig. 3(a), we can see that a long burst occurs at \( G_{K_{Ca}} \approx 4.3 \). An enlarged diagram of Fig. 3(a) around this point is shown in Fig. 3(b). The vertical axis shows an averaged number of spikes in 200 bursts. Figure 4 indicates waveforms of the membrane potentials of the axon for the SC and LC. At \( G_{K_{Ca}} = 4.368 \) we observed persistent normal bursts as shown in Fig. 4(a). Decreasing the value of \( G_{K_{Ca}} \), a long burst suddenly appeared among the normal burst. Further decreasing \( G_{K_{Ca}} \) triggered increase the appearance of a long burst and the alternate firing pattern of the normal and long bursts was observed between \( G_{K_{Ca}} \approx 4.359 \) and 4.364. Finally, persistent long bursts as shown in Fig. 4(b) appeared for \( G_{K_{Ca}} < 4.359 \).

However, we could not observe the second burst of the axon for the LC in one burst for the SC. We investigated burst activity for changing other parameters’ values and
we determined that $G_{KCa}$ for the LC is important to reproduce the experimental results. Figure 5 shows waveforms of membrane potentials for the SC and LC axon. Comparing it with Fig. 2(b) we could reproduce the experimental waveforms. Thus, we conclude that one of causes of generating alternans is $G_{KCa}$; decreasing this conductance triggers alternans. Note that the time scale was changed to $\tau = \omega t$, where $\omega = 2.0$ and $\omega = 4.0$ in Fig. 4(a) and, Figs. 4(b) and 5, respectively. Thus, we have to obtain these figures at the same time scale as one of our open problems.

4. Conclusion

We investigated the parameter dependence of the alternans in a mathematical crustacean cardiac model. 91 parameter planes of conductances of all ionic currents were carefully checked. As a result, we determined that the conductance of calcium-dependent potassium current for the small cell is a key to generating alternans which had been shown in an experiment on a hermit club. Thus, we can reproduce the waveforms of the alternans by changing the value of the conductances. We are now trying to clarify bifurcation structure in the parameter plane. It is said that crustacean hearts and human hearts are fundamentally the same in terms of morphology and physiology [12, 13]; thus, our result could be applicable to the human heart system.

Recently, it is reported that calcium dynamics is important for the occurrence of the alternans [14, 15, 16, 17, 18, 19, 20, 21]. Thus, studying the detailed model for calcium dynamics is also one of our open problems.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 15K00405.

References


Discordant Alternans and Codimension-three Bifurcations in Coupled Luo-Rudy Models

Hiroyuki Kitajima and Toru Yazawa

Abstract—Electrical alternans is the alternating amplitude from beat to beat in the action potential of the cardiac cell. Spatial pattern of the alternans in the heart is classified into two types: concordant and discordant alternans. The former and latter indicate spatially uniform and nonuniform distribution of the alternans. In this study, we investigate bifurcations for a system of coupled two Luo-Rudy models. As a result, we clarify the bifurcational mechanism of generating discordant alternans. Moreover, we determine existence of codimension-three bifurcations related to generation of discordant alternans.

1. Introduction

Electrical alternans is a beat to beat alternation in the action potential duration or amplitude of the cardiac cell [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13]. Alternans is classified into eight or more kinds in [2]. Occurrence of the voltage or calcium-driven cellular alternans is explained by using action potential duration (APD) restitution, calcium dynamics, or voltage-calcium coupling, which is the dynamics in cellular level. On the other hand, the concordant and discordant alternans occur in tissue level. The former and latter corresponds to an in-phase and anti-phase synchronized solution in a system of coupled cells. The discordant alternans is related to QRS alternans. The former and latter indicate spatially uniform and symmetry-breaking appear [22, 23, 24, 25, 26]. Moreover, the intersection of the two period-doubling bifurcations occurs in a parameter plane, which is referred as a codimension-three bifurcation, and the pitchfork bifurcations of a higher periodic solution appear [27].

Using the numerical bifurcation analysis method [28], we determined that the symmetry-preserving and symmetry-breaking period-doubling bifurcations generate the concordant and discordant alternans, respectively. Moreover, we found the parameter region of coexistence of two kinds of alternans. Biexcitability was reported [29, 30, 31], but these are for wave propagation in a two-dimensionally coupled system. In this study, we showed that the biexcitability is observed in a simpler system. We consider that the study of a small number of coupled cells is prototype to study the whole network because some groups composed of synchronized cells can be treated as one cell.

2. Preliminaries

3. System Equations

We use the LR model [20] for the sake of simplicity. The period of the external force (usually called “BCL”: basic cycle length) is assumed to be 380 [ms]. The membrane potentials \( V_1 \) and \( V_2 \) of coupled two LR models with the synaptic external input is described by

\[
C \frac{dV_1}{dt} = -(I_1 + I_{syn} + G_s(V_1 - V_2)),
\]

\[
C \frac{dV_2}{dt} = -(I_2 + I_{syn} + G_s(V_2 - V_1)),
\]

and the type and the dynamics for ionic currents \( I_s \) are given in Appendix. The synaptic current \( I_{syn} \) from the large cell to the muscle cell is given by

\[
I_{syn} = G_{syn}(V - V_{syn})s(t^*),
\]

where \( G_{syn} \) is the maximum synaptic conductance, \( V_{syn} \) is the reversal potential, and \( s(t^*) \) is given by

\[
s(t^*) = \frac{\tau_1}{\tau_2 - \tau_1} \left( -\exp\left( -\frac{t^*}{\tau_1} \right) + \exp\left( -\frac{t^*}{\tau_2} \right) \right),
\]

where \( \tau_1 \) and \( \tau_2 \) are the raise and decay time of the synapse.
the experimental data [32]. \( t^* \) is the time that is reset at every \( nT \) (\( n \) is a natural number, and \( T \) is the BCL). We check the periodicity of the trajectory by using the state variables at every \( nT \). The values of the parameters related with the synapse are fixed as \( G_{\text{syn}} = 4.0 \) and \( V_{\text{syn}} = -29.0 \).

### 3.1. Symmetry

We assume that a system equation is described by

\[
\frac{dx}{dt} = f(x), \quad x \in \mathbb{R}^n
\]  
(5)

A system is called invariant with respect to \( g \) of a group \( G \)
(or \( G \)-equivariant) if

\[
g f(x) = f(g x), \forall g \in G.
\]  
(6)

The fixed-point subspace \( X^G \subset \mathbb{R}^n \) is defined as

\[
X^G = \{ x \in \mathbb{R}^n : g x = x, \forall g \in G \}.
\]  
(7)

The set \( X^G \) is a linear subspace of \( \mathbb{R}^n \). This subspace is an invariant set of Eq. (5) [26, 33].

Considering our system (\( n = 16 \)), the group \( G \) is formed by

\[
G = \{ g \}, \quad g = \begin{bmatrix} O & I_8 \\ I_8 & 0 \end{bmatrix}
\]  
(8)

where \( O \) and \( I_8 \) is zero and identity matrix, respectively. The fixed-point subspace \( X^G \) is given by

\[
X^G = \{ x \in \mathbb{R}^{16} : g x = x \}.
\]  
(9)

We introduce the definitions of symmetric periodic solutions [26].

**Definition 3.1** A periodic solution \( \psi(t) \) is called in-phase if \( g \psi(t) = \psi(t) \) for all \( t \in \mathbb{R} \).

**Definition 3.2** A periodic solution \( \psi(t) \) with (minimal) period \( T_S \) is called anti-phase if

\[
g \psi(t) = \psi(t + T_S / 2)
\]

for all \( t \in \mathbb{R} \).

### 4. Results

We show a bifurcation diagram in Fig. 1. The negative value of \( G_t \) is no meaning in physiological sense, however, it is important from the viewpoint of the dynamical system because of existence of the intersection of double period-doubling bifurcations. Black thick lines denoted by \( I_1 \) and \( I_3 \) present symmetry-preserving period-doubling bifurcations, which means that an in-phase 2-periodic solution (Fig. 2(a)) is generated from an in-phase solution. Period-doubling bifurcation \( I_2 \) is symmetry-breaking which generates an anti-phase 2-periodic solution (Fig. 2(b)). The intersection points of these bifurcations (closed circles in Fig. 1) are called a codimension-three bifurcation. It is known that pitchfork bifurcations of 2-periodic solutions appear from these points, those are \( D^1_2 \) to \( D^2_2 \). We explain the detailed bifurcation structure around this codimension-three bifurcation point using Fig. 3.

Figure 3 is a schematic diagram along curve \( l \) in Fig. 1. At the starting point of \( l \) we observe stable in-phase 1-periodic solution which corresponds to a normal state in the cardiac system. Stable 2-periodic solutions synchronized at anti-phase are generated at \( \odot \). This solution (red curve in Fig. 3) meets pitchfork bifurcation \( D^2_2 \) and becomes unstable at \( \odot \). It disappears at \( \ominus \). The unstable in-phase 1-periodic solution (black dashed line) also meets period-doubling bifurcation at \( \odot \) and unstable 2-periodic solutions are generated. The stability turns to stable at \( \odot \) by pitchfork bifurcation \( D^3_2 \). Thus, stable in-phase and anti-phase 2-periodic solutions coexist between \( \odot \) and \( \ominus \). This region is shown as overlapping of green and hatched pattern in Fig. 1.

Figure 4 shows the difference between solutions denoted by red and green curves in Fig. 3 using the plane (\( V_1, V_2 \)). The diagonal line (\( V_1 = V_2 \)) represents the invariant subspace \( X^G \) given by Eq. (9). The trajectory of anti-phase (Fig. 4(a)) is symmetry with respect to the operation \( g \) in Eq. (8). On the other hand, we obtain another solution by the operation \( g \) for the solution without symmetry (Fig. 4(b)).

### 5. Conclusion

We investigated a system of coupled Luo-Rudy models. We introduced symmetries for the system and its solutions. Using them, we clarified period-doubling bifurcations into two types: symmetry-preserving and breaking. The concordant and discordant alternans are generated by the symmetry-preserving and breaking period-doubling bifurcations, respectively. We determined the parameter region of coexistence of these two alternans.

### Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 15K00405. We thank Prof. T. Yoshinaga of Tokushima University for providing his powerful bifurcation analysis tools.

### References


Detailed explanation of these equations is written in [20].
Figure 1: Bifurcation diagram for coupled LR models. Curves $I$ and $D$ indicate period-doubling and pitchfork bifurcations, respectively. We observe in-phase 1-periodic, anti-phase 2-periodic, and in-phase 2-periodic solutions in gray, green, and hatched regions, respectively.

Figure 2: Waveforms of membrane potentials of two 2-periodic solutions at $G_V = 0.1$ and $[K_0] = 7.2$.

(a) In-phase (two waveforms are identical)

(b) Anti-phase ($V_1$: red, $V_2$: blue)

Figure 3: Schematic bifurcation diagram along curve $l$ in Fig. 1. Solid and dashed curves indicate stable and unstable solutions, respectively. Black and blue curves represent in-phase solution. Red is used for anti-phase solution. Solution without symmetry are shown by green.

Figure 4: Trajectories in $(V_1, V_2)$ plane for two 2-periodic solutions. Arrows indicate the direction of trajectory. (a) one trajectory at $G_V = 0.1$ and $[K_0] = 7.2$, (b) two trajectories denoted black and red at $G_V = 0.1$ and $[K_0] = 6.9$.
Secure Communication via Cluster Synchronization of Chaotic Systems

Zekeriya Sarı† and Serkan Günel‡

† The Graduate School of Natural and Applied Sciences, Department of Electrical and Electronics Engineering, Dokuz Eylül University, 35160 Buca, Izmir, Turkey
‡ Faculty of Engineering, Department of Electrical and Electronics Engineering, Dokuz Eylül University, 35160 Buca, Izmir, Turkey
Email: zekeriya.sari@deu.edu.tr, serkan.gunel@deu.edu.tr

Abstract—In this paper, conceptual design of a secure communication system based on the cluster synchronization of chaotic systems has been proposed. Dividing the network into two sub-networks as transmitter and receiver, the symbols to be transmitted are signified by the cluster modes of the corresponding node groups. Numerical simulations have been presented to illustrate the basic concept.

1. Introduction

The discovery of synchronization of two or more chaotic systems without changing their dynamical characteristics (strange attractors peculiar to chaotic systems, sensitivity to initial conditions, etc.) and direct observations of this phenomenon in physical and biological systems have led to careful investigation of the synchronization of the nonlinear dynamical systems over the past decades [1]. It is well known that due to their sensitive dependence on initial conditions, the chaotic systems are capable of producing unpredictable outputs, characteristically. The idea of using them in secure data transmission has been studied extensively [5, 6, 7, 8].

The problem of under what conditions the interacting dynamical systems are able to synchronize has much more variety of results than expected at a first glance [2, 3, 4]. Depending on the network topology and coupling strength, all the nodes can synchronize to a common (periodic, quasi-periodic or chaotic) dynamical behavior resulting in full synchronization. The nodes of the network can also synchronize in groups forming clusters but there can be no synchronization in between the groups resulting in cluster synchronization [2, 3, 4]. Each cluster can have a different dynamical behavior or can loose synchronization totally.

In this paper, conceptual design of a secure communication system via the cluster synchronization of the chaotic systems has been proposed. In the proposed system, one or more controllable parameters allow cluster synchronization of the different node groups. The symbols to be transmitted are represented by the cluster synchronization modes. The paper is organized as follows: In section 2, design of arbitrary clusters in networks of chaotic systems has been summarized. Section 3 presents the main idea of using cluster synchronization of the chaotic systems to design secure communication systems. The concept has been illustrated through an example in Section 4.

2. Arbitrary Clusters of Networks of Chaotic Systems

Consider the network given as,

\[ \dot{x}_i = f(x_i) + \sum_{j=1}^{N} e_{ij}(t)c_{ij}P_{x_j}, \quad i = 1, 2, \ldots, N \]  

consisting of identical chaotic oscillators where \( x_i \in \mathbb{R}^d \) is the state vector of the \( i^{th} \) node, \( f : \mathbb{R}^d \rightarrow \mathbb{R}^d \) is the vector function defining individual node dynamics, \( e_{ij}(t) \in \mathbb{R} \) is the coupling strength at time \( t \) between the \( i^{th} \) and \( j^{th} \) nodes, \( N \) is the number of the nodes. The diagonal matrix \( P = diag(p_1, p_2, \ldots, p_d) \in \mathbb{R}^{d \times d} \), where \( p_k > 0 \) for \( k = 1, \ldots, s \) and \( p_k = 0 \) for \( k = s+1, \ldots, d \) determines by which state variables of the nodes are coupled. The symmetric matrix with zero row sums \( C = [c_{ij}] \in \mathbb{R}^{N \times N} \) determines the network topology, i.e. \( c_{ij} \neq 0 \) if there exists coupling between the \( i^{th} \) and \( j^{th} \) node and \( c_{ij} = 0 \), otherwise.

Following the formalism in [4], assume that the network is divided into \( n \) clusters as \( G_1 = \{1, \ldots, m_1\}, \ldots, G_n = \{\sum_{k=1}^{n-1} m_k + 1, \ldots, \sum_{k=1}^n m_k = N\} \), where \( m_j \) is the number of nodes in the \( j^{th} \) cluster and \( m_1 \leq \ldots \leq m_n \) can be assumed without any loss of generality. ‘\( i \sim j \)’ denotes that the \( i^{th} \) and \( j^{th} \) nodes are in the same cluster. \( G_k \) is the number of nodes that are in the same cluster with \( k^{th} \) node.

The design of arbitrary clusters and assurance of the stability of them require solution of two problems consequently; constructing the coupling matrix \( C \) and adjusting the coupling strength among the nodes [4].

Assuming that the individual system \( \dot{x}_i = f(x_i) \) is eventually dissipative, which implies there exists compact sets \( B_i = \{x_i \mid \|x_i\| \leq b_1\} \) attracting all the trajectories outside of it, then it can be shown that the coupled network in (1) is eventually dissipative [4].

Considering the case \( e_{ij}(t) = e_{i1}, \quad i, j = 1 \ldots, N, \quad \forall t \) and denoting the error between the \( i^{th} \) and \( j^{th} \) node as \( e_{ij} \)

\[ e_{ij} = x_j - x_i, \quad i, j = 1, \ldots, N, \quad i \sim j \]  

the error dynamics is governed by,
\[ \dot{e}_{ij} = f(x_j) - f(x_i) + \epsilon_1 \sum_{k=1}^{N} (c_{jk} Pe_{jk} - c_{ik} Pe_{ik}) \]  \hspace{1cm} (3)

where \( i, j = 1, \ldots, N, \ i \sim j \). Let \( Df \) denote the Jacobian matrix of \( f \), then (3) can be written as,

\[ \dot{e}_{ij} = \left[ \int_{0}^{1} Df(\beta x_j + (1 - \beta)x_i) d\beta \right] e_{ij} + \epsilon_1 \sum_{k=1}^{N} (c_{jk} Pe_{jk} - c_{ik} Pe_{ik}) \hspace{1cm} (4) \]

Adding and subtracting the term \( Ae_{ij} \) to (4) to avoid instabilities caused by the positive eigenvalues of \( Df \), where \( A = \text{diag}(a_1, \ldots, a_d) \), \( a_k > 0 \) for \( k = 1, \ldots, s \) and \( a_k = 0 \) for \( k = s + 1, \ldots, d \) (recall that \( s \) is the number of nonzero elements in diagonal matrix \( P \)), we have,

\[ \dot{e}_{ij} = \left[ \int_{0}^{1} Df(\beta x_j + (1 - \beta)x_i) d\beta - A \right] e_{ij} + \epsilon_1 \sum_{k=1}^{N} (c_{jk} Pe_{jk} - c_{ik} Pe_{ik}) \hspace{1cm} (5) \]

Considering the auxiliary system,

\[ \dot{e}_{ij} = \left[ \int_{0}^{1} Df(\beta x_j + (1 - \beta)x_i) d\beta - A \right] e_{ij} \hspace{1cm} (6) \]

we assume Lyapunov functions of the form,

\[ W_{ij} = \frac{1}{2} e_{ij}^T H e_{ij}, \ i = 1, \ldots, N, \ i \sim j \]

where \( H = \text{diag}(h_1, \ldots, h_s, H_1) \), \( h_1 > 0, \ldots, h_s > 0 \) and \( H_1 \in \mathbb{R}^{(d-s)x(d-s)} \) is a positive definite matrix, whose time derivative along the trajectories of (6) is negative, i.e.

\[ \dot{W}_{ij} = e_{ij}^T H \left[ \int_{0}^{1} Df(\beta x_j + (1 - \beta)x_i) d\beta - A \right] e_{ij} < 0 \]

\[ \text{if } i = 1, \ldots, N, \ i \sim j, \ e_{ij} \neq 0 \hspace{1cm} (8) \]

**Theorem 2.1.** Under the eventual dissipativeness of (1) and assumption (8) with the coupling matrix,

\[ C = \begin{bmatrix}
3C_{11} & C_{12} & 0 & 0 & \cdots & 0 & 0 \\
C_{21} & 5C_{22} & C_{23} & 0 & \cdots & 0 & 0 \\
0 & C_{32} & 5C_{33} & \ddots & \cdots & 0 & 0 \\
0 & \vdots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 5C_{s-1,s-1} & C_{s-1,n} \\
0 & 0 & 0 & \cdots & C_{n-1,n} & 3C_{n,n} 
\end{bmatrix} \]

(9)

where \( C_{ii} \in \mathbb{R}^{m_i \times m_i} \), \( C_{ii} = [c_{jk}] \), \( c_{jk} = 1 \) for \( j \neq k \) and \( c_{jj} = -\sum_{k=1}^{m_i} c_{jk} \). \( C_{ij} \in \mathbb{R}^{m_i \times m_j} \). \( C_{i+1,j} = C_{j+1,i} = (C_{ij}, 0) \) if \( m_i < m_{i+1} \) or \( C_{i+1,j} = C_{j+1,i} = C_{ii} \) if \( m_i = m_{i+1} \), the cluster synchronization invariant manifold \( M = \{ x_1 = \ldots x_m, \ldots, x_{m+1} = \ldots = x_N \} \) is globally asymptotically stable, if the following inequality holds:

\[ \epsilon_1 \sum_{i=1}^{N} \sum_{j=i}^{N} \bar{G}_{ij} e_{ij}^T H P e_{ij} \geq \epsilon_1 \sum_{i=1}^{N} \sum_{j=i}^{N} m_i e_{ij}^T H A e_{ji} \hspace{1cm} (10) \]

**Proof.** See [4] for the proof. \( \Box \)

**Corollary 2.1.** Since \( \bar{G}_{ij} = \ldots = \bar{G}_{m_1} = m_1 \leq \ldots \leq \bar{G}_R = m_n \), we have,

\[ \epsilon_1 \sum_{i=1}^{N} \sum_{j=i}^{N} m_i e_{ij}^T H P e_{ij} \geq \epsilon_1 \sum_{i=1}^{N} \sum_{j=i}^{N} m_i e_{ij}^T H A e_{ji} \hspace{1cm} (11) \]

Then, if

\[ \epsilon_1 \sum_{i=1}^{N} \sum_{j=i}^{N} m_i e_{ij}^T H P e_{ij} \geq \epsilon_1 \sum_{i=1}^{N} \sum_{j=i}^{N} m_i e_{ij}^T H A e_{ji} \hspace{1cm} (12) \]

holds, which implies,

\[ \epsilon_1 \geq \frac{1}{m_1 \max_{1 \leq k < s \leq p} a_{h}} \hspace{1cm} (13) \]

the Theorem (2.1) is satisfied.

### 3. Secure Communication via Cluster Synchronization

In an attempt to use cluster synchronization in networks of oscillators in a communication system in which the information to be sent is encoded into symbols \( s_i, \ i = 0, \ldots, M-1 \), consider an arbitrary network of \( N \) identical oscillators that can be partitioned into \( n \) clusters. The network evolves with the dynamics in (1). All the oscillators are in their chaotic regime. The whole network is divided into two sub-networks as transmitter and receiver in such a way that the nodes connecting the transmitter side and the receiver side are not in the same cluster. Since the change of cluster mode in the transmitter side also changes the dynamics in the receiver side, the symbols to be transmitted are represented by the absence or the presence of the cluster synchronization of the corresponding node groups. The coupling strengths in the transmitter side are adjusted to assure the cluster mode corresponding to the symbol to be transmitted.

Figure 1a shows the case when the coupling strengths in the transmitter side are adjusted so that the symbol, say, \( s_1 \) is to be transmitted. Note, since all the nodes are in their chaotic regime, and the nodes connecting the transmitter side and receiver side are not in the same cluster, the transmitted signals through the channel are always chaotic and cannot be resolved without the knowledge of the transmitter and the receiver internal topologies. Figure 1b shows the case when the couplings in the transmitter side is adjusted so that another symbol, say, \( s_2 \) is to be transmitted. The spatiotemporal behavior of the network changes when different symbols are to be transmitted.
4. An Example for Binary Communication

Consider the network of identical Lorenz oscillators given in Figure 2 with the network dynamics in (1). The Lorenz system defined by,

\[
\begin{align*}
\dot{x}^{(1)} &= \sigma (x^{(2)} - x^{(1)}) \\
\dot{x}^{(2)} &= x^{(1)} (r - x^{(3)}) - x^{(2)} \\
\dot{x}^{(3)} &= x^{(1)} x^{(2)} - b x^{(3)}
\end{align*}
\]

is chaotic for the parameter values \(\sigma = 10.0, \ b = 8.0/3.0, \ r = 28.0\) [4]. All the oscillators in the network are coupled to each other through their \(x^{(1)}\) states, i.e. \(P = \text{diag}(1,0,0)\).

Consider that the network is to be used in a binary communication system where the information to be sent is encoded into symbols 0 and 1. Assume that the symbols 0 and 1 are signified by the spatiotemporal behavior in Figure 2a and Figure 2b, respectively. Note that for the symbol 1 transmission the clusters \(G_1 = \{1,2\}\), \(G_2 = \{3,4\}\) are formed while for the symbol 0 transmission and must be a non-zero value such that all the oscillators in the network are not synchronous.

Assuming that the transmitted signals through the channel are disturbed by independent additive white Gaussian noise, the network in Figure 2 has been numerically integrated with Euler-Maruyama method for 500 seconds with a step size of 0.005 seconds, signal-to-noise ratio of 20 dB, bit rate of 0.1 bps and \(\epsilon_1 = 10.0, \ \epsilon_0 = 1.0\). Figure 3a shows pulse modulated waveform of the information signal. Figures 3b, 3c and 3d show the time waveforms of the errors between the nodes. Figure 3b and 3d shows the existence of \(G_1 = \{1,2\}\), \(G_2 = \{3,4\}\) clusters during symbol 1 transmission and their disappearance during symbol 0 transmission. Figure 3c shows that since node 1 and node 2 are not in the same cluster, there is no change in the temporal behavior of \(e_{23}(t)\), which is one of the error signals through the channel, during any symbol transmission and \(e_{23}(t)\) is always chaotic. Figure 3e shows the spectrogram of \(e_{23}(t)\) in logarithmic scale plotted with the pulse modulated waveform of the information signal and it can be seen that there is no apparent correlation between the time-frequency properties of \(e_{23}(t)\) and the information to be transmitted, which implies no information can be extracted from the time-frequency properties of the signals transmitted through the channel by a third party, directly, although the system needs to be checked toughly against known attacks.

\[
\mathbf{C}(t) = \begin{bmatrix}
-3u(t) & 3u(t) & -\epsilon_1 & \epsilon_1 \\
3u(t) & -3u(t) & \epsilon_1 & -\epsilon_1 \\
-\epsilon_1 & \epsilon_1 & -3\epsilon_1 & 3\epsilon_1 \\
\epsilon_1 & -\epsilon_1 & 3\epsilon_1 & -3\epsilon_1
\end{bmatrix}
\]

Here, \(u(t) = (1-s_j)\epsilon_0 + s_j\epsilon_1\) which is the pulse modulated waveform of the information signal where \(s_j\) is the symbol to be transmitted at time \(t\).

Note that during the symbol 1 transmission \(u(t) = \epsilon_1\) and (13) must be satisfied for predefined clusters. During symbol 0 transmission, \(u(t) = \epsilon_0\) and must be a non-zero value such that all the oscillators in the network are not synchronous.

Figure 1: An arbitrary network of identical oscillators. The oscillators are illustrated by circles and the oscillators with the same gray level are in the same cluster. Figure 1a and Figure 1b show the cluster mode for the symbol \(s_1\) and \(s_2\), respectively.

Figure 2: A network of identical Lorenz oscillators for a binary communication system. The oscillators are illustrated by circles and the oscillators with the same gray level are in the same cluster. Figure 2a and Figure 2b show the cluster mode for the symbol 0 and 1, respectively.

\([\epsilon_j(t)c_j]\) we have a time varying coupling matrix for the network in Figure 2,

\[
\mathbf{C}(t) = \begin{bmatrix}
-3u(t) & 3u(t) & -\epsilon_1 & \epsilon_1 \\
3u(t) & -3u(t) & \epsilon_1 & -\epsilon_1 \\
-\epsilon_1 & \epsilon_1 & -3\epsilon_1 & 3\epsilon_1 \\
\epsilon_1 & -\epsilon_1 & 3\epsilon_1 & -3\epsilon_1
\end{bmatrix}
\]
5. Conclusion

A conceptual design of a secure communication system based on the cluster synchronization of chaotic systems has been proposed. Numerical simulations including time-frequency properties of the designed system has been presented through an example to illustrate the basic concept. Obtaining the bit error rate of the system, comparing the designed system with the conventional communication systems, investigation of different network topologies and unidirectional coupling are among further research.

References


Desynchronization induced by highly heterogeneous strengths of inhibitory current on electronic neuron

Hisashi Kada† and Isao T. Tokuda‡

†Department of Mechanical Engineering, Ritsumeikan University
1-1-1 Noji-higashi, Kusatsu-shi, Shiga 525-8577, Japan,
Email: [rt001084@ed, isao@fc] .ritsumei.ac.jp

Abstract—In neural networks research, neural information is assumed to be generally represented by a spatio-temporal firing pattern of neurons. When the neurons generate synchronized spike trains with each other, their capability of information representation is severely limited, because all the neurons represent merely a single temporal pattern. On the other hand, when the neuronal activities become desynchronization, the network may exhibit much more variations of their firing patterns, increasing the information processing capability. The aim of the our study is to consider a way how to avoid such synchronized neuronal firings that degrade the information processing capability. First, as a typical situation to induce neuronal synchrony, we introduce common inputs to a feed-forward neural network. Then, to suppress the synchrony, heterogeneity is introduced to inhibitory inputs. Using an electric circuit, we verify that the heterogeneous inputs indeed give rise to asynchronous firings among the electronic neurons.

1. Introduction

The asynchronous irregular activity of the neurons in cortical networks can be spontaneously maintained ongoing firings without external stimuli, which is observed in vitro [1] and in vivo [2]. The various information processing, as a asynchronous irregular state of the cortex, including sensory perception [3], memory [4], signal processing [5] and transmissions [6] in neural networks is reported. To elucidate the underlying mechanism for the cortical network to sustain the spontaneous asynchronous state, the random network with balanced excitatory and inhibitory inputs realizes the asynchronous firing in theoretical studies [7]. In the experiment, entropy is maximum under the asynchronous irregular firings by the balanced inputs [8]. Moreover, the asynchronous firings generate various spatio-temporal patterns of a neuronal activity in the network [9]. The spatio-temporal patterns are brain operations that generate adaptive external inputs. We consider that the asynchronous firing takes information processing capability. When the neural activities synchronized with each other, their capability of information representation is severely limited, because all the neurons represent merely a single temporal pattern.

To avoid such neural activities that degrade the information processing capability, we firstly explore underlying mechanism of neural synchronization. The synchronous firings were observed when the neurons were injected by common inputs with each other [10, 11]. The our study showed that inhibitory common inputs give rise to synchronous firings a feed-forward model with leaky integrate-and-fire neurons based on the physiological experiments [12]. In addition, highly heterogeneous inhibitory post-synaptic potentials (IPSPs) was proposed to suppress their strong synchronization [12]. The introduction of heterogeneous IPSP amplitudes can generate uncorrelated inputs and then the firing pattern can be desynchronized [12].

This paper is to implement a electronic circuit that gives rise to asynchronous firings to avoid synchronous firings induced by inhibitory common inputs. First, using electronic circuit with common inputs, we reproduce a typical situation to induce neuronal synchrony. Next, to suppress the synchrony, we adjust inhibitory inputs to heterogeneity from homogeneity. We verify that asynchronous firings among the electronic neurons can be induced by the heterogeneous inputs.

2. Method

2.1. Electronic neuron circuit

The axon-Hillcock circuit in the electronic neuron circuit is an analog circuit originally proposed by Mead (Fig.1) [13]. The axon-Hillcock circuit imitates the simple dynamic of a neuron with leaky integrate-and-fire model. The input current $I_i(t)$ is stored linearly into the membrane potential $v_i(t)$ of $i$th electronic neuron on the capacitor ($C_1$, $1\mu F$) until the membrane potential exceeds threshold $V_{thr}$. The stored membrane potential $v_i(t)$ is released by n-MOS transistor ($M_1$) and the voltage $V_{thi}$ [14]. The voltage $V_{thi}$ is set to 1.2 V. When the membrane potential $v_i(t)$ reaches threshold $V_{thr}$, the output voltage $V_{out}$ quickly change from 0 to $V_{dd}$ and the electronic neuron generate the spike. The spike can be detected by the sigmoid function, which consists of n- and p-MOS transistors ($M_2$-$M_5$) and capacitor ($C_2$, $47\mu F$). Then a gate voltage of n-MOS transistor ($M_6$) is high by the on-state of the output voltage $V_{out}$. The n-MOS transistor ($M_6$) comes into an on-state from an off-state by the high gate voltage, and then the membrane potential $v(t)$ with a full charge is released to ground and the output volt-
age $V_{\text{out}}$ swings back to 0 V. When the membrane potential $v(t)$ is the reset potential, the n-MOS transistor ($M_6$) return to an off-state from an on-state and the cycle repeats.

The membrane potential in neurons ($v(t)$) is measured by the oscilloscope (Keysight: DSOX2104A). The function generator of arbitrary current (Keysight: N6784A) can inject into the electronic neurons.

$$V_{\text{out}} = \sum_{i=1}^{N} I_i(t) - V_{\text{mem}}$$

Here, $I_i(t)$ is the input current to the electronic neuron and $V_{\text{mem}}$ is the membrane potential. The input currents $I_i(t)$ on the electronic circuit are hard to set, because we adjust a number of coupling strengths more than neurons and we prepare many electronic neurons in first layer. Here, the input current is set to the ensembles of synaptic weights and spikes generated by all neurons in the first layer (Fig. 2B). Concretely, the current is combined with positive and negative pulse waves with a width of 0.01 sec for excitatory and inhibitory inputs, respectively. The input current $I_i(t)$ injected into output neurons is given by

$$I_i(t) = W_{i,E} \sum_k \delta(t-t_{E,k}) + W_{i,I}(t) \sum_k \delta(t-t_{I,k}) + \sum_c \delta(t-t_{I,c}),$$

where the indices E and I are for excitatory and inhibitory neurons, respectively. $\delta(t)$ represents the delta function, $W_{i,E}$ and $W_{i,I}$ are the strengths of excitatory and inhibitory inputs and b-th spike from the neurons in first layer, respectively. Excitatory inputs generate independent Poisson spike trains (red in Fig.2), whereas inhibitory inputs in the first layer generate independent Poisson spike trains (blue in Fig.2) and shared Poisson spike trains (green in Fig.2). We conduct experiments for 20 sec using the 10 electronic neurons and the 10 kinds of input currents.

![Figure 1: Circuit diagram of electronic neuron](image)

Figure 1: Circuit diagram of electronic neuron: The parameters on electronic neuron are set to $C_1 = 1\, \mu F$, $C_2 = 47\, \mu F$, $V_{\text{pw}} = 1.6\, V$, $V_{\text{th}} = 1.2\, V$, and $V_{\text{ih}} = 5\, V$, respectively.

2.2. Input current on electronic neuron

In neural network, it is a generally considered that recurrent networks are composed of two types neurons (excitatory and inhibitory neurons). To simplify recurrent network, we set a feed-forward network that has excitatory and inhibitory neurons in a first layer and 10 neurons in a second layer (Fig. 2A). The neurons in the second layer generate synchronous firings by the input current $I_i(t)$ including inhibitory common inputs [11]. The common input current depends on the spikes and synaptic weights of excitatory and inhibitory neurons in first layer.

The input currents $I_i(t)$ on the electronic circuit are hard to set, because we adjust a number of coupling strengths more than neurons and we prepare many electronic neurons in first layer. Here, the input current is set to the ensembles of synaptic weights and spikes generated by all neurons in the first layer (Fig. 2B). Concretely, the current is combined with positive and negative pulse waves with a width of 0.01 sec for excitatory and inhibitory inputs, respectively. The input current $I_i(t)$ injected into output neurons is given by

$$I_i(t) = W_{i,E} \sum_k \delta(t-t_{E,k}) + W_{i,I}(t) \sum_k \delta(t-t_{I,k}) + \sum_c \delta(t-t_{I,c}),$$

where the indices E and I are for excitatory and inhibitory neurons, respectively. $\delta(t)$ represents the delta function, $W_{i,E}$ and $W_{i,I}$ are the strengths of excitatory and inhibitory inputs and b-th spike from the neurons in first layer, respectively. Excitatory inputs generate independent Poisson spike trains (red in Fig.2), whereas inhibitory inputs in the first layer generate independent Poisson spike trains (blue in Fig.2) and shared Poisson spike trains (green in Fig.2). We conduct experiments for 20 sec using the 10 electronic neurons and the 10 kinds of input currents.

![Figure 2: Schematic imagination of input currents on the electronic neuron](image)

Figure 2: Schematic imagination of input currents on the electronic neuron. (A) Imagination of a feed-forward network with two layers. In the network, Excitatory (triangle) and inhibitory neurons (circle) in the first layer are connected to 10 neurons in a second layer. Excitatory neurons in the first layer generate independent Poisson spike trains (red), whereas inhibitory neurons in the first layer generate independent Poisson spike trains (blue) and shared Poisson spike trains (green). (B) Input current implemented on the electronic neuron corresponding to (A).

2.3. Description of synchronous firing

The level of synchronous firing in the output neurons, we evaluate the Cross-correlogram (CCG). We calculate CCG as a histogram of inter-spike intervals for all pairs among electronic neurons. The time lag was set to the range between -5 sec and 5 sec with an increment of 1 sec. If the neural activity can be synchronized with each other, the CCG appears with a sharp peak at zero time lag, whereas asynchronous firings indicate the existence of the CCG with a flat structure.
Figure 3: Effect of inhibitory common inputs on synchronous firings of the electronic neuron: (A) Raster plots corresponding to the case that no common inputs (c = 0) were injected. Spikes of electronic neurons are indicated. (B) Raster plots corresponding to the case that the ratio to shared inputs c is set to 0.77. (C,D) The CCG histograms corresponding to (A,B), respectively.

3. Results

3.1. Experiment with synchronization induced by common inputs of inhibitory current

For the desynchronization, we first set synchronous firings on the electronic neurons by inhibitory common inputs. Then, we introduce a network mechanism that reduces the correlated inhibitory inputs.

The parameters of input current $I(t)$ are defined as strengths and frequencies of excitatory and inhibitory inputs. The mean strengths of excitatory and inhibitory inputs are constant for $W_{x,E} = W_x = 1$ mA and $W_{y,I}(t) = W_I = -1.25$ mA, respectively. Moreover, $R_E = 8$ Hz indicate the frequency of independent Poisson spike trains for excitatory inputs, whereas inhibitory inputs are classified into independent Poisson spike trains $R_f$ and shared Poisson spike trains $R_{f,I}$. The frequency of inhibitory inputs $R_I$ is defined as the summation of each frequency $R_{f,I}$ and $R_f$, and $R_f = 4$ Hz is always constant in all experiments. To constrain correlated spike trains of inhibitory neurons, we introduce the ratio to shared spikes $c = R_{f,I}/R_I$.

The raster plots under the two states with c = 0 and c = 0.77 are shown in Fig. 3. As comparing correlated inputs c = 0.77 (Fig. 3B) with uncorrelated inputs c = 0 (Fig. 3A), synchronous firing with c = 0.77 occurs than that with c = 0. As shown in CCG, the peak of CCG for uncorrelated inputs (Fig. 3C) is less than correlated inputs (Fig. 3D). The result is that the synchronous firings can be reproduced by the correlated inhibitory inputs on the electronic neurons.

3.2. Experiment with desynchronized firings induced by heterogeneous inputs of inhibitory current

In the previous subsection, we showed that synchronous firings are generated by high common inputs. We conduct the experiment with desynchronization under the c = 0.77, which induce synchronous firings in electronic neurons. To suppress the synchronous firings, the strengths of inhibitory inputs $W_{y,I}(t) = W_I$ are distributed such that lognormal distribution has high heterogeneity (Fig. 4);

$$p(w_i) = \frac{1}{\sqrt{2\pi}\sigma w_i} \exp\left[\frac{-(\log w_i - \mu)^2}{2\sigma^2}\right]$$

where we set the mean of the distribution $\overline{W_I}$ and the value $\sigma$, respectively [15]. Once their parameters $\overline{W_I}$ and $\sigma$ are decided, we can obtain the parameter $\mu = \log(\overline{W_I}) - \sigma^2/2$. The mean of the distribution $\overline{W_I} = -1.25$ mA keeps the same value among experiments.

As shown in the raster plots with current strengths under the constant $\sigma = 0$ (Fig. 3B), the synchronous firings still remains, whereas the firing pattern is asynchronous under lognormal distribution with $\sigma = 3$ (Fig. 5A). We also understand that the peak of CCG exists under the constant (Fig. 3D), while the CCG disappears with a sharp peak under the lognormal distribution (Fig. 5B). The implies is that the highly heterogeneous strengths of inhibitory inputs have an impact on desynchronization.

4. Discussion

We implemented the electronic circuit that is mimics the neural activity proposed by Mead [13] and we introduced an input current injected into a neuron in output layer to the highly correlated inhibitory current. As increasing the ratio to shared inputs, we confirmed that the neurons in output layer are synchronized with each other. To break the highly correlated inhibitory current, we changed to heterogeneous strengths of inhibitory current. The input current can reduce the correlated inhibitory inputs and then the asynchronous firings can be observed on the electronic circuit.
Figure 5: Effect of standard deviation of inhibitory strengths on the electronic neuron. The proportion to shared inputs $c$ is set to 0.77. (A) Raster plot represents spikes of electronic neurons under he inhibitory inputs distributed as lognormal distribution with $\sigma = 3$. (B) The CCG histograms corresponding to (A).

We consider that the asynchronous firings on the electronic neuron are more information capacity than the synchronous firings. It is reported that the spontaneous asynchronous firing activity gives rise to the maximum information capacity in the physiological experiments [8]. The pathological synchronous firings that may relate to epileptic seizure [16] reduce information capacity because all the neurons represent merely a single neural activity. In the our future, we discuss the relationship between the pathological synchronous firings and information capacity on electronic neurons.

References


Synchronization phenomena of piecewise-linear oscillators coupled by a hysteresis element

Makoto Sudou† and Tadashi Tsubone‡

†Nagaoka University of Technology
1603–1, Kamitomiokamachi,Nagaoka, Niigata, 940–2401 Japan
Email: s111051@stn.nagaokaut.ac.jp

Abstract—In this study, we consider two-dimensional piecewise-linear oscillators with a diode coupled by a hysteresis element. Coexistence phenomena of in-phase and an anti-phase synchronization are confirmed. We have examined parameter conditions where coexistence phenomena can be observed. Typical behavior are confirmed in laboratory.

1. Introduction

In nature and engineering fields, we observe often a synchronization phenomena which is one of the interesting non linear phenomena. For example, synchronization chirp of frog and synchronization of the neural network to perform the information processing of the human brain are well known[1]. However, because of complexity of large-scale network, it is often difficult to analyze the phenomena. Therefore, coupled systems of oscillators by using electrical circuits are good model for considering the synchronization phenomena. The van der Pol oscillator is one of the famous nonlinear oscillators, and it have been studied by many researches[2]. It is well-known that in-phase and anti-phase synchronization are observed in the van der Pol oscillators coupled by a resister and a native resister, respectively. On the other hand, in the case of the oscillators coupled by a inductor, a capacitor, a negative conductor and a diode. If the diode is assumed to operate as ideal switch, the circuit equation is degenerate for one-dimensional when diode is on[5]. We confirm coexistence phenomena of in-phase and anti-phase synchronization in two-dimensional piecewise-linear oscillators with a diode coupled by hysteresis element in laboratory experiments. We have examined parameter conditions for coexistence phenomena of in-phase and an anti-phase synchronization by using the bifurcation diagram.

2. Piecewise-linear oscillators with a diode

Figure1 shows a circuit diagram. The oscillator is two-dimensional piecewise linear oscillators with a diode. The element $-g$ is a negative conductor. We represent the $v-i$ characteristics of a negative conductor and an idealized diode (see Fig.2) . $i_D(v)$ presents the current of the negative conductor, and $i_D(v)$ presents the current through the diode. The diode is assumed an ideal switch with the threshold voltage of $E$. The circuit dynamics is represented by one-dimensional equation when the diode is on. In addition, the threshold voltages of the diode can be varied by connecting some diodes in series. The
circuit dynamics can be represented by the following twodimensional piecewise-linear differential equation:

Diode OFF

\[
\begin{align*}
L \frac{di}{dt} &= v \\
C \frac{dv}{dt} &= -i - i_D(v),
\end{align*}
\]

(1)

Diode ON

\[
\begin{align*}
L \frac{di}{dt} &= E \\
v &= E(i_D = i - i_D(E)).
\end{align*}
\]

(2)

If Diode is OFF \((v < E)\), diode becomes ON when \(v\) reaches \(E\).

If Diode is ON \((i_D > 0)\), diode becomes OFF when \(i_D\) reaches 0.

By using normalized variable and parameter as follows:
\[
\tau = \frac{1}{\sqrt{LC}} t, \quad \delta = \frac{g}{E} \sqrt{\frac{L}{C}}, \quad x = \frac{1}{E} \sqrt{\frac{L}{C}} i,
\]
\[
y = \frac{x}{E}, \quad f_D = \frac{1}{E} \sqrt{\frac{L}{C}} i_D,
\]
the following normalized equation is obtained:

Diode OFF

\[
\begin{align*}
\frac{dx}{d\tau} &= y \\
\frac{dy}{d\tau} &= -x + 2\delta y,
\end{align*}
\]

(3)

Diode ON

\[
\begin{align*}
\frac{dx}{d\tau} &= 1 \\
y &= 1.(f_D = x - 2\delta).
\end{align*}
\]

(4)

Each region is connected to another region by the following transition conditions:

Figure 3 shows the numerical results and laboratory experiments. It was confirmed that these results match qualitatively.

Figure 4 shows a circuit diagram. This circuit is two-dimensional piecewise-linear oscillators with a diode coupled by a hysteresis element. VCCS with hysteresis characteristic as shown in Figure 5. \(H\) is switched from 1 to -1 if \(v_1 - v_2\) reaches to the threshold \(-v_{th}\) and \(H\) is switched from -1 to 1 if \(v_1 - v_2\) reaches to \(v_{th}\).
The circuit dynamics can be represented as follows:

**Diode OFF**

\[
\begin{align*}
L \frac{dI_1}{dt} &= v_1 \\
C \frac{dv_1}{dt} &= -i_1 - i_G(v_1) - i_c \\
L \frac{dI_2}{dt} &= v_2 \\
C \frac{dv_2}{dt} &= -i_2 - i_G(v_2) + i_c,
\end{align*}
\]

**Diode ON**

\[
\begin{align*}
L \frac{dI_1}{dt} &= E \\
v_1 &= E(i_D = i_1 - i_G(E) + i_c) \\
L \frac{dI_2}{dt} &= E \\
v_2 &= E(i_D = i_2 - i_G(E) - i_c).
\end{align*}
\]

By using following variables and parameters:

\[
\begin{align*}
\tau &= \frac{1}{\sqrt{LC}}, \delta = \frac{q}{T}, \gamma = \frac{1}{T} \sqrt{E_i}, \alpha = \frac{1}{T} \sqrt{E_i}, \\
y_1 = \frac{y_1}{E}, y_2 = \frac{y_2}{E}, \gamma = \frac{1}{E} \sqrt{E_i}, \\
f_{D1} = \frac{1}{T} \sqrt{E_i D_1}, f_{D2} = \frac{1}{T} \sqrt{E_i D_2},
\end{align*}
\]

the normalized equation is obtained:

**Diode OFF**

\[
\begin{align*}
\frac{dx_1}{d\tau} &= y_1 \\
\frac{dy_1}{d\tau} &= -x_1 + 2\delta y_1 - \gamma h(y_1 - y_2) \\
\frac{dx_2}{d\tau} &= y_2 \\
\frac{dy_2}{d\tau} &= -x_2 + 2\delta y_2 + \gamma h(y_1 - y_2)
\end{align*}
\]

**Diode ON**

\[
\begin{align*}
\frac{dx_1}{d\tau} &= 1 \\
y_1 &= 1(f_{D1} = x_1 - 2\delta + \gamma h(y_1 - y_2)) \\
\frac{dx_2}{d\tau} &= 1 \\
y_2 &= 1(f_{D2} = x_2 - 2\delta - \gamma h(y_1 - y_2)).
\end{align*}
\]

Figure 6 shows laboratory experiments and numerical results under same parameter conditions. Figure 6(a) shows in-phase synchronization, and Figure 6(b) shows anti-phase synchronization. In these result, we confirm coexistence phenomenon of in-phase and anti-phase synchronization in this system. Further, from the results represented in Fig.6(c) and (d), it was identified the co-existence phenomenon of in-phase and anti-phase synchronization in the circuit experiment. These behaviors are qualitatively consistent with behavior of van der Pol oscillators coupled by a inductor.
4. Parameter regions of in-phase and anti-phase synchronization

In order to derive the parameter regions of in-phase and anti-phase synchronization, we introduce correlation coefficients $c$ with regard to $y_1$ and $y_2$.

\[
c = \frac{\sum_{k=0}^{N} (x_{1k} - \overline{x_1})(x_{2k} - \overline{x_2})}{\sqrt{\sum_{k=0}^{N} (x_{1k} - \overline{x_1})^2 \sum_{k=0}^{N} (x_{2k} - \overline{x_2})^2}} \tag{9}
\]

Figure 7 shows the bifurcation diagram of in-phase and anti-phase synchronization. $\Delta \alpha$ is parameter mismatch of oscillator 1 and 2. $\Delta \theta h$ is hysteresis width of hysteresis element.

5. Conclusion

In this paper, we consider two-dimensional piecewise-linear oscillators with a diode coupled by hysteresis element. Firstly, we confirmed the behavior of the two-dimensional oscillator including inductor, capacitor, negative conductor and a diode from the numerical results and laboratory experiments. These results match qualitatively. Next, we investigated the behavior of piecewise-linear oscillators coupled by a hysteresis element. As a result, we confirmed coexistence phenomena of in-phase and anti-phase synchronization. Finally, we illustrated parameter conditions where coexistence phenomena of in-phase and an anti-phase synchronization. We will investigate parameter conditions where coexistence phenomena of in-phase and an anti-phase synchronization in more detail.

References


Stability Analysis of Phase-Inversion Waves in Coupled Piecewise-Constant Oscillators As a Ladder

Ryohei Hirota† and Tadashi Tsubone‡

†Nagaoka University of Technology
1603-1, Kamitomioka, Nagaoka, Niigata, 940-2188, Japan
Email: s133165@stn.nagaokaut.ac.jp
‡Nagaoka University of Technology
Email: tsubone@vos.nagaokaut.ac.jp

Abstract—In this study, we consider phase-inversion waves in coupled piecewise-constant oscillators as a ladder. A nonlinear phenomenon called phase-inversion wave is a kind of wave propagation phenomena, and the phase states of the waves are propagated to next oscillator in succession. In this paper, we analyzed the stability of the phase-inversion waves using the largest Lyapunov exponent. As a result, we found out that phase-inversion waves is unstable and chaos, because the largest Lyapunov exponent is positive value.

1. Introduction

There are many reports for analysis of synchronization phenomena of coupled oscillators [1–5]. Suzuki and Tsubone have confirmed that piecewise-constant oscillators coupled by hysteresis elements exhibit co-existence of in-phase and anti-phase synchronization [1]. They also analyzed the stability of the system by Lyapunov exponents. Yamauchi, Nishio, and Ushida have discovered wave propagation phenomena called phase-inversion waves of coupled van der Pol oscillators [2, 3]. The phase-inversion waves are wave propagation phenomena, and the phase states of the waves are propagated to next oscillator in succession. It is very important to analyze the phenomena, because it is similar to propagation phenomena of electrical information in an axial fiber of nervous systems. However, if nonlinearity of van der Pol oscillators are strong, the analysis often becomes hard. The simulation requires high calculation cost. When we carry out Lyapunov analysis for high dimensional system, sometimes very long time is needed. Therefore, confirmation of phase-inversion waves in rigorous sense and detailed stability analysis of the systems which generate the phase-inversion waves have not been discussed. Accordingly, we consider piecewise-constant oscillators. The oscillators are simple systems and the analysis is relatively easy. The systems have piecewise-constant vector fields, and the solutions are piecewise-linear. Hence, we have only to focus on the borders of switching of the vector fields, we can determine the rigorous solutions [1]. Using the calculation method [1], we can derive the rigorous solution with low calculation cost.

In this paper, we show phase-inversion waves of coupled three piecewise-constant oscillators. We also analyze the stability of the phase-inversion waves in our system by the largest Lyapunov exponent. As a result, we found out that phase-inversion waves is chaos, because the largest Lyapunov exponent is positive value. It can be obtained rigorously from computer-aided analyzing procedure by using rigorous solutions.

2. Circuit model

2.1. A Piecewise-Constant Oscillator

Figure 1 shows a circuit model of a piecewise-constant oscillator.

Fig. 1 Cicuit model of a piecewise-constant oscillator.

\[
\begin{align*}
\frac{dv_1}{dt} &= I_1 \cdot H(v_1) + I_3 \cdot \text{sgn}(v_2); \\
\frac{dv_2}{dt} &= I_2 \cdot H(v_1).
\end{align*}
\]

(a)VCCS with hysteresis characteristic.
(b)VCCS with Signum-like function.

Fig. 2 Symbols and nonlinear characteristics of VCCSs.

The circuit equations of the system are described as follows.

\[
\begin{align*}
\frac{dv_1}{dt} &= I_1 \cdot H(v_1) + I_3 \cdot \text{sgn}(v_2), \\
\frac{dv_2}{dt} &= I_2 \cdot H(v_1),
\end{align*}
\]
where $I_1$, $I_2$ and $I_3$ are absolute values of output currents of hysteresis or sgn Voltage Controlled Current Sources (VCCSs). We consider the following conditions.

$$I_2 = -I_3, \quad I_1 \cdot I_2 < 0. \quad (2)$$

The conditions (2) guarantees non-constrained behaviors. $H(v_{th})$ and $\operatorname{sgn}(v_{th})$ are hysteresis and signum characteristic respectively, as shown in Fig. 2. We use following dimensionless variables and parameters

$$\tau = \frac{I_2}{C \cdot v_{th}}, \quad x = \frac{v_1}{v_{th}}, \quad y = \frac{v_2}{v_{th}}, \quad \alpha = -\frac{I_1}{I_2}. \quad (3)$$

Then, we can rewrite the circuit Eq. (1) as following dynamics,

$$\begin{align*}
\dot{x} &= -\alpha h(x) - \operatorname{sgn}(y) \\
\dot{y} &= h(x),
\end{align*} \quad (4)$$

where “.” denote differentiation by normalized time $\tau$, $h(X)$ shows normalized hysteresis. If $X$ reaches 1, the output switches from -1 to 1, and if $X$ reaches -1, output switches from 1 to -1. The system has only one parameter $\alpha$. In order to oscillate, we consider the following conditions [1].

$$0 < \alpha < 1. \quad (5)$$

Figure 3 shows a rigorous solution and the corresponding laboratory measurement.

![Rigorous solution and laboratory measurement](image)

(a) Rigorous solution. (b) Laboratory measurement.

Fig. 3 Rigorous solution and laboratory measurement of a piecewise-constant oscillator.

### 2.2. Piecewise Constant Oscillators Coupled by hysteresis element as a Ladder

We consider piecewise-constant oscillators coupled by hysteresis elements as a ladder. Figure 4 shows circuit model of the coupled piecewise-constant oscillators. This system has following dynamics.

[First oscillator] ($m = 1$)

$$\begin{align*}
\dot{x}_1 &= -\alpha h(x_1) - \operatorname{sgn}(y_1) - \gamma h(x_1 - x_2) \\
\dot{y}_1 &= h(x_1),
\end{align*} \quad (6)$$

[Middle oscillator] ($2 \leq m \leq N - 1$)

$$\begin{align*}
\dot{x}_m &= -\alpha h(x_m) - \operatorname{sgn}(y_m) \\
&\quad - \gamma h(x_m - x_{m+1}) + \gamma h(x_{m-1} - x_m) \\
\dot{y}_m &= h(x_m),
\end{align*} \quad (7)$$

The system has two parameters, $\alpha$ and $\gamma$. $N$ is number of oscillators, and $\gamma$ is a coupling parameter. In this paper, we discuss the case of $N = 3$. In our previous study, we observed phase-inversion waves in this system [5]. Figure 5 shows a rigorous solution and the corresponding laboratory measurement.

![Laboratory measurement](image)

(a) Rigorous solution. (b) Laboratory measurement.

Fig. 5 Phase-inversion waves in rigorous solution and laboratory measurement.

Figure 5 represents the difference between output of adjacent oscillators. Small waves are in-phase, and large waves are anti-phase synchronization. We can observe changing phase state, namely phase-inversion waves.
3. Calculation of rigorous solution and Lyapunov exponent

In piecewise-constant system, we can obtain the rigorous solution directly by noting borders of vector fields [1]. The borders are points that $h$ or sgn switches. For easy to explain, we introduce the algorithm for solution be given a piecewise-constant oscillator in this section. A trajectory starts initial point $x_0 = (x_0, y_0)$, and goes straight forward to border $E$ in accordance with vector field $a(i)$ in Table 1. A time to reach the border $\tau$ is obtained by

$$\tau_x = \frac{E_x(i) - x}{x}, \quad \tau_y = \frac{E_y(i) - y}{y},$$

(9)

where $\tau$ is positive and minimum value. We calculate $x_{k+1}$ using $\tau$ and $x_k$.

$$x_{k+1} = x_k + a(i_k) \cdot \tau.$$  

(10)

In the same manner, we can also obtain $x_2$, $x_3$... and derive rigorous solution, like Fig. 6.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$h(x)$</th>
<th>$\text{sgn}(y)$</th>
<th>$a(i)$</th>
<th>$E_x(i)$</th>
<th>$E_y(i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>(-$\alpha - 1, 1$)</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>-1</td>
<td>(-$\alpha + 1, 1$)</td>
<td>-1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-1</td>
<td>1</td>
<td>($\alpha - 1, -1$)</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>-1</td>
<td>-1</td>
<td>($\alpha + 1, -1$)</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1: Local vector fields and borders for $i$.

In Fig. 8, phase-inversion wave happens and the largest Lyapunov exponent is $\lambda_1 = 7.3421 \times 10^{-3} > 0$. On the other hand, we can observe in-phase or almost in-phase synchronization phenomena in white regions. Figure 7 is a two-parameter bifurcation diagram. The vertical axis is $\alpha$ and the horizontal axis is $\gamma$. In the figure, black and gray regions show $\lambda_1 > 0$. White regions show $\lambda_1 = 0$. Reds show $\lambda_1 < 0$. We defined that $|\lambda_1| < 10^{-7}$ is $\lambda_1 = 0$ [7].

Fig. 7 Two-parameter bifurcation diagram in $\gamma - \alpha$ plane.

In gray regions, we can observe phase-inversion waves. On the other hand, we can observe in-phase or almost in-phase synchronization phenomena in white regions.

Next, we compared waveforms in gray and white regions. Figure 8 and 9 show output differences in $(\gamma, \alpha) = (-0.03, 0.2)$ and $(\gamma, \alpha) = (-0.03, 0.3)$.

**Time 500[τ/div]**

![Time 500[τ/div]](image)

In Fig. 8, phase-inversion wave happens and the largest Lyapunov exponent is $\lambda_1 = 7.3421 \times 10^{-3} > 0$. On the other hand, we can observe in-phase or almost in-phase synchronization phenomena in white regions.

Next, we compared waveforms in gray and white regions. Figure 8 and 9 show output differences in $(\gamma, \alpha) = (-0.03, 0.2)$ and $(\gamma, \alpha) = (-0.03, 0.3)$.

![Time 500[τ/div]](image)

In Fig. 8, phase-inversion wave happens and the largest Lyapunov exponent is $\lambda_1 = 7.3421 \times 10^{-3} > 0$. On the other hand, we can observe in-phase or almost in-phase synchronization phenomena in white regions.

Next, we compared waveforms in gray and white regions. Figure 8 and 9 show output differences in $(\gamma, \alpha) = (-0.03, 0.2)$ and $(\gamma, \alpha) = (-0.03, 0.3)$.

![Time 500[τ/div]](image)

In Fig. 8, phase-inversion wave happens and the largest Lyapunov exponent is $\lambda_1 = 7.3421 \times 10^{-3} > 0$. On the other hand, we can observe in-phase or almost in-phase synchronization phenomena in white regions.

Next, we compared waveforms in gray and white regions. Figure 8 and 9 show output differences in $(\gamma, \alpha) = (-0.03, 0.2)$ and $(\gamma, \alpha) = (-0.03, 0.3)$.
other, in-phase synchronization happens and the largest Lyapunov exponent is $\lambda_1 = 1.2437 \times 10^{-9} \approx 0$ in Fig. 9. Figure 9 looks like periodic motion. However, when we confirm waveforms of longtime, we found that this is non-periodic motion. In Fig. 10, we can see that waveforms change subtly.

We show output differences in other parameters. Figure 11 and 12 show output differences in $(\gamma, \alpha) = (-0.02, 0.2)$ and $(\gamma, \alpha) = (-0.01, 0.1)$. The largest Lyapunov exponents are both positive value. Thus, we can observe another type phase-inversion waves, but we can not divide these phenomena with the largest exponent.

5. Conclusion

In this paper, we considered phase-inversion waves of three piecewise-constant oscillators coupled by hysteresis elements as a ladder. We carried out stability of phase-inversion waves with the largest Lyapunov exponent. As a result, we found that phase-inversion waves are unstable and chaos because the largest Lyapunov exponent is positive value. Our future tasks are calculating the 2nd and 3rd largest Lyapunov exponents and analyzing in more detail by using these numbers.

References

Interval Estimation of Coupling Delay Time from Phase Time Series

Boris Bezruchko,†‡ Elena Sidak,† and Dmitry Smirnov‡

†Faculty of Nano- and Biomedical Technologies, Saratov State University
83 Astrakhanskaya Str., Saratov 410012, Russia
‡Saratov Branch, V.A. Kotel’nikov Institute of Radioengineering and Electronics RAS
38 Zelyonaya Str., 410019 Saratov, Russia
Email: bezruchkobp@gmail.com, sidakev@gmail.com, smirnovda@yandex.ru

Abstract – We consider interval estimation of a time delay in coupling between oscillatory systems from observed time series. It is shown that asymptotic estimates, based on an empiric model in the form of first-order phase oscillators and maximum likelihood formalism, can lead to false conclusions about the value of the delay in two cases: (i) for nonlinear low-dimensional systems whose phases are well-defined but considerable amplitude fluctuations make phase description of the dynamics insufficient, (ii) for systems whose phases are not well-defined due to large amplitude fluctuations. We suggest an empirical criterion for diagnosis of such problematic situations and develop a modified estimator assuring low probability of false conclusions in those situations. Efficiency of the suggested estimator is demonstrated for benchmark systems with different dynamical properties, including stochastic and deterministically chaotic oscillators. An application of the approach to an analysis of large-scale climate processes is presented.

1. Introduction

Phase approximation is widely used to describe dynamics of oscillatory systems in nonlinear dynamics and oscillation theory [1,2]. Due to reduction of model dimensionality and retaining essential dynamical properties, phase description appears an efficient approach to a series of problems including the study of synchronization conditions [2,3]. Due to high sensitivity of the phase variable to external influences, phase dynamics analysis is used to detect couplings between oscillators from time series [4-6] which appears useful in different fields including neurophysiology and climatology (e.g. [7] and references therein). Within such an analysis, it is important to have tools for coupling delay estimation [8], e.g. to estimate signal propagation time in a complex medium. When dealing with short time series (several dozens of basic periods) typical in practice, it is important to get not only the value of the delay (a point estimate), but also a justified estimate of its uncertainty (an interval estimate). Several recent studies is devoted to the development of such tools [9,10]: the estimation methods are based on fitting empirical models in the form of coupled first-order phase oscillators where future phases are determined by the current phase values and external noises generated independently of the current phase values. Such a model is strictly justified for self-sustained oscillators which individually exhibit limit cycles, while couplings and noises slightly perturb these cycles [1]. Under violation of those conditions, including intensive noises leading to strong amplitude fluctuations or chaotic regimes of low-dimensional nonlinear systems, the phase description is insufficient and the first-order phase oscillators appear quite a rough approximation. Still, one may expect that the phase model-based coupling delay estimators are sometimes applicable even in such complicated cases, but it is to be checked. Further, for practical applications one needs a criterion to recognize situations where the existing delay estimators are erroneous. Then, one also needs modified estimators applicable to such problematic situations. The questions of revealing problematic situations, finding a criterion for their recognition, and developing a modified estimator are studied in this work. Section 2 describes an existing interval estimator of coupling delay. Section 3 presents a typical situation, where that interval estimator can be erroneous, and introduces a modified estimator assuring low pre-defined error probability. Section 4 gives an application of the approach to real-world data about large-scale climate processes El-Nino/Southern Oscillation and North Atlantic Oscillation. Conclusions are given in Section 5.

2. Asymptotic Interval Estimator

According to the technique developed in Refs. [8-10], one computes phases of the observed signals \( x_1(t) \) and \( x_2(t) \) (e.g. using the analytic signal construction [2] as in the examples below) and gets the time series \( \{\phi_1(t_1)\ldots\phi_1(t_N)\} \) and \( \{\phi_2(t_1)\ldots\phi_2(t_N)\} \) from observed signals \( \{x_1(t_1)\ldots x_1(t_N)\} \) and \( \{x_2(t_1)\ldots x_2(t_N)\} \), where \( t_j = j\Delta t, \Delta t \) is sampling interval, \( N \) is time series length. Then, one fits the phase dynamics model whose form comes from the fact that the phase dynamics of weakly perturbed periodic self-sustained oscillators yields to the differential equations of the first-order stochastic phase oscillators

\[
\frac{d\phi_k(t)}{dt} = \omega_k + G_k(\phi_k(t), \phi_j(t - \Delta_j^{(k)}) + \xi_k(t)),
\]

\[ k, j = 1,2, j \neq k, \]
where independent white noises possess covariance functions \( \langle \xi_k(t) \xi_k(t') \rangle = D_k \delta(t-t') \), functions \( G_k \) determine both individual phase nonlinearity of the oscillators and their couplings, and \( \Delta^*_j \) are coupling delay times. In time series analysis, on fits stochastic difference equations which correspond to the equations (1) integrated over an interval of the finite length \( \tau \) (a parameter of the method):

\[
\phi_k(t + \tau) - \phi_k(t) = F_k(\phi_k(t), \phi_j(t - \Delta_j \rightarrow k)) + \xi_k(t),
\]

\( k, j = 1, 2, j \neq k \)

where \( \Delta_j \rightarrow k \) is a trial time delay, \( F_k \) is a low-order trigonometric polynomial, whose coefficients are determined via minimization of the mean squared model error

\[
\delta^2 = \left( \langle \phi_k(t_j + \tau) - \phi_k(t_j) \rangle^2 - \langle \phi_k(t_j) \rangle^2 \right)^{1/2}.
\]

An achieved minimal value \( \delta^2(\Delta_j \rightarrow k) \) is then minimized over \( \Delta_j \rightarrow k \) : its minimum point is \( \hat{\Delta}_j \rightarrow k = \arg \min \delta^2(\Delta_j \rightarrow k) \). An unbiased coupling delay estimate for the system (1) then reads \( \hat{\Delta}^*_{j \rightarrow k} = \hat{\Delta}_j \rightarrow k + \tau/2 \).

An asymptotic maximum likelihood estimator of its variance is given by

\[
\hat{\sigma}^2_{\Delta_j \rightarrow k} = \frac{2 \hat{\sigma}^2}{N'} \left( \frac{\partial^2 \delta^2(\Delta_j \rightarrow k)}{\partial \Delta_j \rightarrow k} \right)_{\Delta_j \rightarrow k = \hat{\Delta}_j \rightarrow k}^{-1},
\]

where \( N' \) is the number of statistically independent values of model residual errors over the time series. It is estimated as \( N' = N/T \), where \( L = \max[T, \tau] \) and \( T \) is the decay time of the autocorrelation function of the model residual errors for the \( k \)th oscillator. The 95% confidence interval for the time delay is then given by

\[
[\hat{\Delta}^*_{j \rightarrow k} - 2 \hat{\sigma}_{\Delta_j \rightarrow k}; \hat{\Delta}^*_{j \rightarrow k} + 2 \hat{\sigma}_{\Delta_j \rightarrow k}] \] and its width is

\[ M = 4 \hat{\sigma}_{\Delta_j \rightarrow k} \]

Efficiency of this interval estimator is shown in Refs. [9,10] for coupled phase oscillators perturbed by white or colored noise and for van der Pol generators. Error results are observed if the dependency \( \delta^2(\Delta_j \rightarrow k) \) does not exhibit a single clear minimum which occurs in case of too small noise level in the driving oscillator or too large noise level in the driven oscillator.

3. Problematic Situations and Modified Estimator

Let us consider low-dimensional nonlinear systems with well-defined phases and possibility of chaotic dynamics where problems with the above estimation technique can be expected due to violation of the assumption about one-dimensional phase dynamics perturbed by external noises (1). Such an example is given by Roessler systems:

\[
\begin{align*}
\dot{x}_1(t) &= -\omega_1 y_1(t) - z_1(t) + \xi_1, \\
\dot{y}_1(t) &= \omega_1 x_1(t) + ay_1(t), \\
\dot{z}_1(t) &= \gamma - z_1(t)(r - x_1(t)), \\
\dot{x}_2(t) &= -\omega_2 y_2(t) - z_2(t) + K(x_1(t - \Delta) - x_2(t)) + \xi_2, \\
\dot{y}_2(t) &= \omega_2 x_2(t) + ay_2(t), \\
\dot{z}_2(t) &= \gamma - z_2(t)(r - x_2(t)),
\end{align*}
\]

(3)

where \( \omega_1 = 1.015, \omega_2 = 0.985 \), \( a = 0.1 \), \( b = 0.1 \), and parameter \( r \) has been varied in a wide range allowing transition from periodic to chaotic regimes via a period-doubling cascade, \( \xi_1, \xi_2 \) are white noises of intensities \( D_{\xi_1,2} \), coupling delay time is \( \Delta_0 = 12 \), \( K \) is coupling coefficient. If one defines phases via the relationships \( x_{1,2} = A_{1,2} \cos \phi_{1,2} \) and \( y_{1,2} = A_{1,2} \sin \phi_{1,2} \), then the phase dynamics of the driven system yields to the equation

\[
\frac{d\phi_2}{dt} = \omega_2 + \frac{z_2(t) \sin \phi_2(t) + (K + a) \sin 2\phi_2(t)}{A_2(t)} - \frac{2A_2(t)}{KA_1(t) \sin (\phi_2(t) - \phi_1(t - \Delta))} - \frac{\hat{\xi}_2(t) \sin \phi_2}{A_2(t)} - \frac{2A_2(t)}{KA_1(t) \sin (\phi_2(t) + \phi_1(t - \Delta))} - \frac{2A_2(t)}{KA_1(t)}.
\]

Even at zero \( D_{\xi_1,2} \), in the system (3), the reduced phase model (1) must include “noise” terms approximating the influence of the amplitude \( A \) and the third coordinate \( z \). Properties of such “efficient phase noises” may be rather non-trivial, especially in chaotic regimes, leading to larger-than-expected errors in the above asymptotic estimates.

In numerical simulations we generated ensembles of 100 time series of the variables \( x_{1,2} \) at each set of parameter values: the integration step in the Euler integration scheme was 0.001, the sampling interval \( \Delta \tau = 0.3 \) (20 data point per a basic period), the length of each time series \( N = 2000 \) (about 100 basic periods) or \( N = 20000 \) (i.e. 1000 periods), the parameter \( r = 1.5 \) (it gives an optimal sensitivity of the method, though the results are weakly sensitive to its values in the range from a quarter to several basic periods). From each pair of time series we computed their phases via Hilbert transform and obtained the above interval estimates of the delay. Then, the number of erroneous estimates (i.e. such that \( \Delta \) does not belong to the interval \( [\hat{\Delta}^*_{j \rightarrow k} - 2 \hat{\sigma}_{\Delta_j \rightarrow k}; \hat{\Delta}^*_{j \rightarrow k} + 2 \hat{\sigma}_{\Delta_j \rightarrow k}] \) was counted and the frequency of the errors \( f_{err} \) was calculated. We say that the estimator works properly if \( f_{err} < 0.1 \), because error probability of 0.05 corresponds to the claimed 0.95 confidence band and finite-ensemble fluctuations of the error frequency distributed according to Bernoulli’s law make an allowable error frequency level somewhat larger. The results of the analysis for the
individually chaotic oscillators with \( r = 10 \) and coupling coefficient \( K = 0.05 \) are shown in Fig.1. Autocorrelation function of the phase model residual errors (Fig.1,a) decays down to a small value (about 0.2) after several oscillations with doubled basic oscillation period due to peculiarities of the Rössler attractor structure. At time series length of 100 basic periods, situations “good” for the asymptotic interval estimator correspond to noise-perturbed regimes with \( \sqrt{D_{r_1}} > 0.6 \) (Fig.1,b, circles). At weaker noises, \( f_{err} \) exceeds the threshold level of 0.1, i.e. the asymptotic estimator becomes unreliable. The difficulty can be diagnosed in practice [9,10] from the absence of a clear minimum on the plot \( s_k^2(\Delta_{j->k}) \) (left inset in Fig.1,b).

The situation becomes even more difficult if longer time series are considered (1000 basic periods, probably less important in practical applications): large error frequency is observed at high noise levels as well (Fig.1,c, circles). It occurs because the point estimator \( \hat{\Delta}_{j->k} \) appears somewhat biased, while the variance estimator \( \hat{\sigma}^2_{\Delta_{j->k}} \) remains correct and becomes small (of the order of nonzero estimator bias) for long time series leading to frequent erroneous estimates. The results are analogous for other forms of couplings (e.g. couplings introduced into the equation for the \( y \)-variable) and for strongly perturbed individually-periodic systems. The cause of the nonzero bias seems to be in the peculiarities of the interaction between the phases and amplitudes and \( z \)-coordinates ignored in the phase model (2). The bias can be different for different nonlinear systems, so that one needs a special criterion for the practical recognition of possible difficulties and a modified estimator to get reliable estimates in such problematic situations.

We suggest to regard an essentially non-quadratic form of the minimum of \( s_k^2(\Delta_{j->k}) \) as a sign of “danger”. Namely, one should be careful in using the asymptotic interval estimator if the minimum is skewed (asymmetric, as in the right inset in Fig.1,b and inset in Fig.1,c), sufficiently deep local minima exist in the vicinity of the global minimum, etc. In such cases, we suggest to go beyond the asymptotic estimators based on approximation of local properties of the \( s_k^2 \) minimum and focus on its global properties which may be expected to be a more robust feature. Namely, we draw a straight line parallel to the abscissa axis at the mean level between maximal and minimal values of \( s_k^2 \) within the range of trial delays considered (e.g. a range from zero to five basic periods used above) and take its leftmost and rightmost cross-section points with the plot \( s_k^2(\Delta_{j->k}) \) as the boundaries of the interval estimator. Such a modified (rough) estimator eliminates high error frequencies in all the above problematic situations (Fig.1,b,c, triangles) at the expense of typically somewhat wider confidence band \( M \) (Fig.1,d,e, triangles). However, it still allows an informative estimation, distinguishing an existing coupling delay from zero in the above examples. The modified estimator may also appear unreliable only if any clearly pronounced minimum on the plots \( s_k^2(\Delta_{j->k}) \) (even a skewed one) is absent (left inset in Fig.1,b) which is easily diagnosed in practice.

The asymptotic estimator encounters the same large-error problems also in cases when the oscillators’ phases are not well-defined due to strong amplitude fluctuations induced either by external random perturbations or internal chaotic dynamics. We showed that in numerical simulations with chaotic Lorenz systems and stochastic linear oscillators. The modified estimator allowed us to avoid frequent erroneous conclusions in all these cases as well.

4. Application to climate processes

Both estimators are applied to an analysis of couplings between El-Nino/Southern Oscillation (ENSO) and North Atlantic Oscillation (NAO) from observational data. These climate processes represent leading modes of interannual climate variability [11]. Influence of ENSO on
NAO was detected in Ref. [12] where a point estimate of the delay time in this influence was obtained. The latter appeared equal to about 20-24 months. The above interval estimators may give additional information about probable range of values of that delay. We have used the following indices reflecting those processes: the leading mode of 500 hPa geopotential height surface for NAO [13] and sea surface temperature in equatorial Pacific (region Nino-3.4: 5°N-5°S, 170°W-120°W) for ENSO [11].

As shown in Ref. [12], the phases defined via complex Morlet wavelet transform [14-16] in the frequency band corresponding to the periods from 24 to 40 months allow us to detect the influence of ENSO to NAO. Here, we applied the above interval estimators to the same phase time series. The results are following: The corrected point estimate of the delay time is 36 months. The asymptotic interval estimate cover the range 29-43 months. However, the plot for the model prediction errors reveals sing that the modified estimator may be more appropriate. The latter gives a wider confidence interval from 8 to 47 months. Still, not approaches detect nonzero coupling delay. However, the modified estimators evidences that it is not necessary to seek for the causes of the 2-year delay. Quite probably, the delay is smaller and consists slightly more than half a year. Such information is of interest for climate science as discussed in Ref. [12].

5. Conclusions

We have studied applicability of the asymptotic interval estimator of the coupling delay time based on phase dynamics modeling, to oscillators with different dynamical properties. We have shown that low-dimensional nonlinear dynamics along with strong amplitude fluctuations can strongly increase probability of erroneous conclusions about the value of the delay. An empirical criterion for diagnosis of such problematic situations is suggested, along with a modified interval estimator based on the rough (global or larger-scale) properties of the phase model residual errors’ minimum. The latter estimator is shown to provide the error probability less than a pre-defined small value for characteristic oscillatory systems with rather different properties of phase dynamics. Hence, it extends possibilities of reliable coupling delay time estimation for a wide range of oscillatory systems in practice. In particular, we have applied the suggested approach to analyze large-scale climate processes from observational data where it confirms a non-zero delay in the influence of El-Nino/Southern Oscillation on the North Atlantic Oscillations.

Acknowledgments

The work is supported by the Russian Science Foundation (grant No. 14-12-00291).

References

Synchronized bifurcation in a two-coupled Izhikevich neuron model

Yuu Miino and Tetsushi Ueta

†Graduate School of Advanced Technology and Science, Tokushima University
2-1, Minamijosanjima-cho, Tokushima, 770–8506
‡Institute of Technology and Science, Tokushima University 2-1, Minamijosanjima-cho, Tokushima 770–8506
Email: nagamonnyuu@gmail.com, ueta@tokushima-u.ac.jp

Abstract—In a coupled Izhikevich neuron model, there are some parameter regions where each neuron arises firing at the same time. We observed discontinuous changes of stability of the solutions in these regions. In this study, we regard these changes as a bifurcation phenomena. We present their examples and condition of the bifurcation. Finally we calculate sets of the bifurcation.

1. Introduction

Izhikevich neuron model[1] can present a lot of general firing patterns, which is observed in real neurons. In addition, the model can keep the costs of calculation lower than the other models. From these reasons, this model is well used in many researches. For example, Tamura[2] proposed the first result of bifurcation analysis for the model. On the other hands, Ito[3] suggested a method of bifurcation analysis for two-coupled Izhikevich neuron model. The study[3] is important for the application such that the neural network. They tried to avoid matching of the two timing. One is the time at firing and the other is the time at the Poincaré map. In this study, we focus on this matching. In coupled Izhikevich neuron model, firings of each neuron synchronize. The phenomenon has not been observed in [3] and any other researches. We found that the phenomenon makes quantitative changes of ω-limit set of the system. That is, the phenomenon can be regarded as a global bifurcation phenomenon. We call this phenomenon as “synchronized bifurcation” in this study. This study shows the characteristics of this study, the method how we calculate the bifurcation sets and the changes how it makes to the ω-limit set of this system.

2. Izhikevich neuron model and its coupled system

The neuron model proposed by Izhikevich[1] is given by

\[
\begin{align*}
\frac{dv}{dt} &= 0.04v^2 + 5v + 140 - u + I \\
\frac{du}{dt} &= a(bv - u)
\end{align*}
\]

where, \(v\) and \(u\) are state variables and \(a, b, I\) and \(\delta\) are parameters. Firing phenomena are realized by following maps:

\[
\begin{align*}
\text{if} \ v &\leq 30, \text{ then } v \mapsto c \\
\text{ } &\mapsto u + d
\end{align*}
\]

where, \(c\) measures a voltage after firing and \(d\) measures strength of the restoration.

Gap-junction two-coupled Izhikevich neuron model is given by

\[
\frac{dv_i}{dt} = f(v_i, a_i, \delta) = \begin{cases} 
0.04v_i^2 + 5v_i + 140 - u_i + I - \delta(v_2 - v_1) & \text{ if } v_1 > 30, \\
0.04v_i^2 + 5v_i + 140 - u_i + I - \delta(v_1 - v_2) & \text{ otherwise}
\end{cases}
\]

where, \(v = (v_1, u_1, v_2, u_2)\) is a vector for state variable and \(\delta\) is strength of the junction. For each neuron, firings arise with the following condition:

\[
\text{if } v_i \leq 30, \text{ then } \begin{cases} 
v_i &\mapsto c \\
u_i &\mapsto u + d
\end{cases}, \ i = 1, 2.
\]

On this study \(a_1\) and \(\delta\) are variable and the other are static.

\[
a_2 = 0.2, b_1 = 0.2, b_2 = 0.2, c = -50, d = 2, I = 10.
\]

Figure 1 shows some time waves of system (3), where, \(t_i\) is the time when \(v_i\) fires. Figure 1(a) shows a stable periodic solution whose \(t_1 > t_2\). Through undergoing the situation \(t_1 < t_2\): synchronized as Fig.1(b), firing order changes as \(t_1 < t_2\). Then stability of the solution becomes unstable. That is, stability of the solution immediately changes since undergoing the synchronized firing.

3. Changing of stability of periodic solution

To evaluate the stability of periodic solutions in this system, the method proposed by Kousaka[4] is strongly effective. The method[4] can solve the bifurcation problem of hybrid system. Hybrid system has digital states (modes) and analog states (states) in its structure. Each mode transits immediately and discontinuously from one to the other(s). These changes called mode transition. Each state evolves by time-continuous or time discrete dynamical system(s).

From the result of previous study[3], let us define the Poincaré section as follows:

\[
\Pi_0 = \{v = (v_1, u_1, v_2, u_2) \in \mathbb{R}^4 | q(v) = v_1 = 0\}.
\]
Figure 1: Time waves of $v_1$(red) and $v_2$(blue) with (a) $a_1 = 0.168, \delta = 0.09$: stable 2-periodic solution, (b) $a_1 = 0.17, \delta = 0.09$: synchronized 2-periodic solution and (c) $a_1 = 0.171, \delta = 0.09$: unstable 2-periodic solution.

Let us consider the solutions starting from $v_0$ on $\Pi_0$ and return to a state on $\Pi_0$ via $m$-times mode transition. $q_k(v) = 0$ is a condition equation for $k$-times mode transition and $\Pi_k = \{ v \mid q_k(v) = 0 \}$ is a manifold expanded by $q_k(v) = 0$. The solution starting from $v_k \in \Pi_k$ at the time $t$ is

$$v_k(t) = \varphi_k(t, v_k, t_k, \lambda),$$

where, $\lambda$ is a certain parameter and

$$v_k(t_k) = v_k = \varphi_k(t_k, v_k, t_k, \lambda).$$

Local map from $\Pi_k$ to $\Pi_{k+1}$ is

$$T_k : \Pi_k \to \Pi_{k+1},$$

$$v_k \mapsto v_{k+1} = T_k(v_k) = \varphi_k(t_{k+1}, v_k, t_k, \lambda).$$

From Eq.(9), the Poincaré map $T$ is expanded as

$$T = T_{m-1} \circ \ldots \circ T_0(v_0).$$

Especially when

$$T^l(v_0) - v_0 = 0,$$  \(11\)

the solution $v_0$ is called as $l$-periodic solution. The Poincaré map $T$ has some trivial factors related to a normal vector of $\Pi_0$. To exclude the factors, let us define local coordinate system $\Sigma \subset \mathbb{R}^{m-1}$ and local coordinate $u = (u_1, u_2, u_1) \in \Sigma$.

$$p^{-1} : \Sigma \to \Pi_0,$$

$$p : \Pi_0 \to \Sigma.$$

The Poincaré map on $\Sigma$ is give by

$$T_\ell : \Sigma \to \Sigma$$

$$u \mapsto p \circ T \circ p^{-1}.$$

On local coordinate system, Eq.(11) is

$$u_0 = T_\ell(u_0) = p \circ T \circ p^{-1}(u_0).$$  \(14\)

The derivative of the Poincaré map $T$ with respect to the initial value $v_0$ is

$$DT = \frac{\partial T_\ell}{\partial u_0} = \frac{\partial p}{\partial v} \frac{\partial T}{\partial v} \frac{\partial p^{-1}}{\partial u_0} = \frac{\partial p}{\partial v} \frac{1}{\partial v} \sum_{k=0}^{m-1} \frac{\partial T_k}{\partial v_k} \frac{\partial p^{-1}}{\partial u_0}. $$

Each derivative is given by

$$\frac{\partial T_k}{\partial v_k} = \begin{bmatrix}
1 - \frac{1}{\partial q_{k+1}} \int \frac{\partial q_{k+1}}{\partial v} \frac{\partial \varphi_k}{\partial v_k}
\end{bmatrix} $$

$$\frac{\partial p}{\partial v} = \begin{bmatrix}
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1
\end{bmatrix},$$

$$\frac{\partial p^{-1}}{\partial u_0} = \begin{bmatrix}
0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}.$$

where, $I$ is a 4 x 4 identity matrix. $\frac{\partial \varphi_k}{\partial v_k}$ is derived by following ordinary differential equation:

$$\frac{d}{dt} \frac{\partial \varphi_k}{\partial v_k} = \frac{\partial f}{\partial v} \frac{\partial \varphi_k}{\partial v_k},$$

with $\frac{\partial \varphi_k}{\partial v_k} = I$.

The characteristic equation is given by

$$\chi(\mu_j) = \det(DT - \mu_j I) = 0, \ j = 1, 2, 3,$$  \(20\)

where, $\mu_j$ is characteristic multiplier, which measures the stability of the Poincaré map $T$. That is, $\mu_j$ can be an index of stability of a periodic solution. When $\forall j, |\mu_j| < 1$, the solution is stable. When $\exists j, |\mu_j| > 1$, the solution is unstable.

Figure 2 shows a root locus that presents how characteristic multiplier changes by undergoing the synchronized firing. Each of Fig.2(a)-(b) denotes that the synchronized firing changes values of all $\mu_j$ immediately and discontinuously. Especially for (a), changes of $\mu_j$ affect the stability of solution since $\mu_j$ goes between the regions where $|\mu_j| > 1$ and $|\mu_j| < 1$ via the synchronized firing. We call this case as synchronized bifurcation.

When we focus on the firing order and the product operation of matrices, the cause of these phenomena is unveiled. $DT$ of the solutions shown on Fig. 1 is expanded as

$$\frac{\partial T}{\partial v_0} = \begin{bmatrix}
\frac{\partial T_3}{\partial v_3} & \frac{\partial T_2}{\partial v_2} & \frac{\partial T_1}{\partial v_1} & \frac{\partial T_0}{\partial v_0}
\end{bmatrix}.$$  \(21\)

For the case Fig.1(a),

$$\begin{bmatrix}
\frac{\partial T_3}{\partial v_3} & \frac{\partial T_2}{\partial v_2} & \frac{\partial T_1}{\partial v_1} & \frac{\partial T_0}{\partial v_0}
\end{bmatrix} = \begin{bmatrix}
0 & 0 & 0 & 0 \\
0.004 & 0.997 & 0 & 0 \\
0.005 & 0 & 1.021 & -0.019 \\
0.005 & 0.001 & 0.9962 & 0
\end{bmatrix}.$$
method is effective. The method needs $DT$ and $\partial i/\partial u_0$ for calculation. This factor is derived by previous method[5].

The local bifurcation sets are obtained by solving Eq. (20) with a condition $\mu_j = 1$.

Figure 3 shows the result of bifurcation analysis and Fig.4 presents phase portraits on each points of Fig.3. At point (a) on Fig.3, there is a stable 1-periodic solution shown on Fig.4(b) at point (b). When parameters are set at point (c), there are no stable 2-periodic solutions and we can observe chaos shown on Fig.4(c). On the region including the point (d), we can observe a 2-periodic solution shown as Fig. 4(d). When seeing $v_1$, $v_1$, the solution has similar structure to the solution on Fig. 4(b). When seeing $v_1$-$v_2$ plane, the solutions are exactly different. This is because the order of firing has changed between these two solutions. The changing is caused by $SF^2$ in Fig. 3.

5. Conclusion

This study investigated the synchronized firing phenomena(SF) observed in 2-coupled Izhikevich neuron model. In our research,

- a drastic changing of index for stability of a periodic solution has arose via SF,
- SF has sometimes made change the stability of a periodic solution: stable to unstable (and vice versa),
- condition of arising SF has been derived,
- a set of parameters where SF arises has been obtained.

For future work, we should try to confirm SF in 3 or more coupled neuron model.

References

Figure 3: Bifurcation diagram with $a_1 \in (0.14, 0.2), \delta \in (0.08, 0.11)$. Red solid curves are set of synchronized firing. Red broken curves are set of synchronized bifurcation. Grey region presents unstable solutions. Blue region presents 1-periodic solutions. Red region presents 2-periodic solutions. Green region presents 4-periodic solution. $G^i$ means tangent bifurcation, $I^i$ means period-doubling bifurcation and $SY^i$ means synchronized bifurcation, respectively from $i$-periodic solutions.

Figure 4: Phase portrait of system (3)–(4) on (top) $v_1 - u_1$ plane and (bottom) $v_1 - v_2$ plane. (red points: Poincaré map, $\delta = 0.1$, (a) 1-periodic solution: $a_1 = 0.155$, (b) 2-periodic solution: $a_1 = 0.163$, (c) chaos: $a_1 = 0.168$, (d) 2-periodic solution: $a_1 = 0.195$)
Almighty Google knows everything! - Big-data and Network Science

Hawoong Jeong †

†Department of Physics, Korea Advanced Institute of Science and Technology

Abstract

Network science is an interdisciplinary academic field which studies complex networks such as engineered networks, information networks, biological networks, and social networks etc. This field has received a major boost caused by the availability of huge network data resources on the Internet. The field draws on theories and methods including graph theory from mathematics, statistical mechanics from physics, data mining and information visualization from computer science, and social structure analysis from sociology to understand the complex systems, the problem to be solved in 21st century. Yet, another research field gaining huge attention nowadays is about big-data. Big-data is defined as “high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.” by Gartner, Inc. This field of research has huge potential for practical applications but it also promises new discovery in science. However, these big-data should be combined and analyzed together to be useful, and in this respect, network science will shed a light on analyzing these big-data in more combined way. In this presentation, I will briefly review what we can do by combining big-data, especially using Google and network science together to study various complex systems such as social network between people, biological networks, and prediction of science and technology trends & even presidential election results etc.

Profile

Prof. Hawoong Jeong is currently KAIST-chair professor at physics department at KAIST, Korea. He got his Ph.D. in physics at Seoul National University, and his research area includes complex systems, statistical computational physics and interdisciplinary science. He published about 100 research papers with more than 15,000 citations in diverse areas including physics, computer science, social science and biology. He got several awards including KAIST best lecturer, KPS research prize, the Scientist of the month award. He has also been selected as Young Scientist at 2012 Summer Davos World Econo Forum.

Education

- 1998: Seoul National University (Ph. D. in Physics)
- 1993: Seoul National University (M.S. in Physics)
- 1991: Seoul National University (B.S. in Physics)
Employment History

- 2001–Present: KAIST, Assistant/Associate/Professor/KAIST-Chair-Professor
- 1998–2001: Univ. of Notre Dame, Post-doc/Research Assistant Professor

Research Interests

- Complex systems
- Structure and dynamics of complex networks
- Bioinformatics
- Molecular dynamic simulation
- Computational methods in statistical physics
- Dynamics of fluctuating interfaces and growing surfaces

Selected Honors & Awards

- 2016 National President’s Commendation (Science and Technology Medal)
- 2014 Hall of Fame: 100 people who will lead Korea after 10 years by Dong-A Newspaper
- 2013 Research Prize by Korean Physical Society (KPS)
- 2012 Young Scientist by Summer Davos World Economic Forum
- 2010 The Scientist of the Month (May 2010) by MOST &KRF
- 2009 Grand prize for excellence in teaching, by KAIST
- 2007 Yong-Bong Prize by Korean Physical Society (KPS)

International Activities

- 2011/7– Associate Editor for BMC Biophysics
- 2010/3– Review Editorial Board of Frontier in Systems Biology
- 2015/7– Chairman of NetSci2016 / Board member of Network Science Society
- 2012/7– Editorial Board of EPJ Data Science
- 2012– International Advisory Committee of STATPHYS25

Selected Publications (out of 100, over 15,000 total citations)

(for full list, see http://stat.kaist.ac.kr/publications.php)

Koopman Operator Theory for Nonlinear Dynamical Systems:
An Introduction with Engineering Applications

Yoshihiko Susuki† and Igor Mezić‡

†School of Electrical and Information Systems, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531, Japan
‡Department of Mechanical Engineering, University of California
Santa Barbara, CA 93106–5070, United States
Email: susuki@eis.osakafu-u.ac.jp, mezic@engineering.ucsb.edu

Abstract—Koopman operator is a composition operator defined for a dynamical system described by nonlinear differential or difference equation. Although the original system is nonlinear and evolves on a finite-dimensional state space, the Koopman operator itself is linear but infinite-dimensional (evolves on a function space). This linear operator captures the full information of the dynamics described by the original nonlinear system. In particular, spectral properties of the Koopman operator play a crucial role in analyzing the original system. In the first part of this presentation, we review the so-called Koopman operator theory for nonlinear dynamical systems, with emphasis on modal decomposition and computation that are direct to wide applications. Then, in the second part, we present a series of applications of the Koopman operator theory to power and energy systems engineering. The applications are established as data-centric methods, namely, how to use massive quantities of data obtained numerically and experimentally, through spectral analysis of the Koopman operator. For the details of this presentation, see our review paper in NOLTA, IEICE (Y. Susuki, I. Mezić, F. Raak, and T. Hikihara, Applied Koopman Operator Theory for Power Systems Technology, vol.7, no.4, October 2016).
Modeling of Effective Heat Diffusion in a Building Atrium via Koopman Mode Decomposition

Yohei Kono†1, Yoshihiko Susuki†2, Mitsunori Hayashida‡3, and Takashi Hikihara†

†Department of Electrical Engineering, Kyoto University
Nishikyo, Kyoto 615-8510, Japan
‡Environmental Business Headquarters, OMRON Corporation
Shiokoji-Horikawa, Shimogyo, Kyoto 600-8530, Japan
Email: y-kono@dove.kuee.kyoto-u.ac.jp

Abstract—We develop a method for modeling of heat transfer dynamics in a building atrium based on observational data on temperature. This paper introduces a two-dimensional heat diffusion equation with an effective diffusion coefficient in order to represent the heat transfer mainly due to the air movement inside an atrium, where the air slowly moves on a length-scale smaller than the distance between rooms. Then we propose a method for identifying the coefficient based on a spatio-temporal oscillatory pattern extracted from the data via Koopman mode decomposition. The calculated coefficient is verified with the characteristic numbers of the air flow in the atrium and its architectural geometry.

1. Introduction

In-building energy dynamics appear on a wide range of scales in both space and time. In the stage of architectural design, lumped-parameter models are used in order to predict coarse-scale thermal dynamics: see e.g. [1]—ranges of more than lengths between rooms. The models have focused on the heat transfer via walls, ceilings, or floors with large time-constant about 100 hours. On the other hand, the heat transfer in a building atrium is on smaller time-scale [2], and thus a short-term change of in-room temperature can propagate between rooms. Therefore, mathematical modeling of the heat transfer in an atrium is required for its evaluation and control.

In this paper, we address the phenomenon of heat transfer mainly due to the air movement inside a practically-used atrium, where the air slowly moves on a length-scale smaller than the distance between rooms. The heat transfer in a building atrium is on smaller time-scale [2], and thus a short-term change of in-room temperature can propagate between rooms. Therefore, mathematical modeling of the heat transfer in an atrium is required for its evaluation and control.

In this paper, we address the phenomenon of heat transfer mainly due to the air movement inside a practically-used atrium, where the air slowly moves on a length-scale smaller than the distance between rooms. The heat transfer is modeled as a two-dimensional heat diffusion equation with an effective diffusion coefficient. Then we propose a method for identifying the coefficient based on a spatio-temporal oscillatory pattern extracted from measurement data via Koopman mode decomposition [3, 4]. The calculated coefficient is verified with the characteristic numbers of fluid flow in the atrium and its architectural geometry.

2. Modeling target in commercial building

This paper focuses on a practically-used atrium in order to delineate the modeling idea. This atrium is located in the main building of OMRON Healthcare Co., Ltd. in Kyoto, Japan. This section provides basic information on the atrium as the modeling object: spatial arrangements and operational conditions of HVAC units, and the measurement of temperature.

Figures 1 and 2 show the cross section of the target building and the outline of a floor. The building consists of six floors, where the eastern sides of the 3rd to 6th floors are provided as offices and the western sides as laboratories. The atrium is located in the center of the building and connected with the office rooms.

Next, we review spatial arrangements and operational conditions of HVAC units. Three HVAC units are placed in the office room for conditioning globally in space. In the below, these units are called H1, H2, and H3. Their outlet ducts are on the ceiling and denoted by the rectangular symbols. The red, green, and blue symbols represent the ducts of H1, H2, and H3, respectively. The HVAC units run on a programed schedule.

Finally, we review the measurement system of temperature in the target building. All the temperature data were...
3. Koopman mode decomposition of measured data

This section applies the Koopman Mode Decomposition (KMD) to the measured data on temperature shown in Fig. 3.

3.1. Outline

KMD is a method for extracting spatio-temporal modes oscillating with single frequencies directly from data [3, 4]. In this paper, we use the Arnoldi-type algorithm [4] to decompose snapshots $X[n] \in \mathbb{R}^M$ ($n = 0, \ldots, N - 1$) with constant time-step $\Delta t$ into the following series:

$$X[n] = \sum_{m=1}^{N-1} \lambda_m^n \tilde{V}_m \quad (n = 0, \ldots, N - 2),$$

where $\tilde{\lambda}_m \in \mathbb{C}$ is Koopman Eigenvalue (KE) and $\tilde{V}_m \in \mathbb{C}^M$ is Koopman Mode (KM).

The sampling period $\Delta t$ is 10 min. Mode pair $[m, m+1]$ that has large values of $|\tilde{\lambda}_m|$ and $|\tilde{V}_m|$ is dominant in the data. In this paper, we adopt mode pair $[5, 6]$ as the dominant one and utilize it to identify the heat transfer in the atrium.

Next, Fig. 4 shows the amplitude and phase of KMs $V_5^A$ and $V_5^H$. The arrows represent the directions of the wavenumbers $k_5^A$ of $V_5^A$ and $k_5^H$ of $V_5^H$. Note that $k_5^A$ and $k_5^H$ are Koopman Mode (KM) directions of the wavenumbers.

Table 1: Koopman mode decomposition of the data measured on July 31, 2014 (Thu.) shown in Fig. 3.

| $[m, m+1]$ | $|\tilde{\lambda}_m|$ | $T_m$ | $|\tilde{V}_m|^2$ | $|\tilde{V}_m|$ |
|-----------|-----------------|------|----------------|-------------|
| [1, 2]    | 1.050           | 2.37 | $7.83 \times 10^{-2}$ | 0.286       |
| [3, 4]    | 1.035           | 4.74 | $5.62 \times 10^{-2}$ | 0.695       |
| [5, 6]    | 1.035           | 3.35 | 0.119          | 0.540       |
| [7, 8]    | 1.033           | 1.16 | $5.52 \times 10^{-2}$ | 0.339       |
| [9, 10]   | 1.033           | 1.86 | $9.31 \times 10^{-2}$ | 0.262       |
| [11, 12]  | 1.026           | 0.547| $2.34 \times 10^{-2}$ | 0.116       |

Here, we formulate spatio-temporal oscillatory pattern of KM as a wave propagation. Let be $m$-th KE defined as

$$\tilde{\lambda}_m = |\tilde{\lambda}_m| \exp(i\omega_m \Delta t),$$

where $i$ is the imaginary unit and $\omega_m$ is the angular frequency. Then the oscillatory pattern of $m$-th KM is represented as follows:

$$|\tilde{\lambda}_m|^n \tilde{V}_m = |\tilde{\lambda}_m|^n \left[ A_{m,1} \exp[i(\omega_m n \Delta t + \theta_{m,1})] \right] \quad (2)$$

where $A_{m,p}$ and $\theta_{m,p}$ stand for modulus (amplitude) and angle (phase) of $p$-th component of KM $V_m$, respectively. Now we represent the observation positions of data by using $r_1, \ldots, r_M \in \mathbb{R}^N (N_d \in \{1, 2, 3\})$ and transposition of vector by the superscript $T$. When the matrix $[r_1^T, \ldots, r_M^T]$ is of full rank, it is possible to uniquely determine $k_m \in \mathbb{R}^N$ satisfying $r_p^T k_m = -\theta_{m,p}$ for $p = 1, \ldots, M$. Then the right-hand side of Eq. (2) is formulated as follows:

$$|\tilde{\lambda}_m|^n \tilde{V}_m = |\tilde{\lambda}_m|^n \left[ A_{m,1} \exp[i(\omega_m n \Delta t - r_1^T k_m)] \right] \quad (3)$$

This shows that the oscillatory pattern of $m$-th KM describes a spatio-temporal wave propagating with the angular frequency $\omega_m$ and wavenumber vector $k_m$.

3.2. Results

This subsection applies KMD to the data shown in Fig. 3. Here, we set $M = 24$ and $N = 55$ (corresponding to 9 hours). In this paper, we decompose KM $\tilde{V}_m$ into the two components $V_m^A \in \mathbb{C}^{12}$ and $V_m^H \in \mathbb{C}^{12}$, where $V_m^A$ stands for the component on the atrium temperature and $V_m^H$ for the outlet temperature.

Table 1 shows the modal information on the data in decreasing order of the modulus $|\tilde{\lambda}_m|$. The eigenperiod is defined as:

$$T_m = \frac{2\pi \Delta t}{||\text{Im}[\ln|\tilde{\lambda}_m|]||}.$$
are the two-dimensional vectors \( N_3 = 2 \) calculated by replacing \( \nabla^H_{m} \) with \( V_5^3 \) and \( V_5^H \) in Eqs. (2) and (3). In Fig. 4(a-1,b-1), the amplitude changes in space, implying that the heat input from HVAC units propagates in the atrium. In Fig. 4(a-2,b-2), the phase also changes in space, and the wavenumber vector \( k_5^H \) is in the opposite direction of \( k_5^3 \). This indicates that the HVAC units operate to decline the heat transfer in the atrium.

4. Identification of heat transfer in atrium

This section formulates the heat transfer in the target atrium as isotropic diffusion equation, namely, effective diffusion. Then, the effective diffusion coefficient of the equation is identified with the KM \( \tilde{V}_5 \) shown in Fig. 4.

4.1. Effective diffusion by small-scale air movement

The target atrium is regarded as the two-dimensional domain because the width of the atrium is much smaller than the length \( x \) and height \( z \) (see Figs. 1 and 2). Here, we introduce the concept of effective diffusion for describing the heat transfer in the two-dimensional domain. It is widely known that the homogenization makes it possible to represent the heat transfer caused by small-scale air movement as isotropic diffusion [5, 6]. In the target atrium, no large-scale laminar flow by HVAC outlets occurs, and hence the air moves on the scale less than the length between rooms. Therefore, by coarse-graining of the thermal dynamics on the length-scale larger than the distance between rooms, we represent the temperature \( T \) at position \( x \) as \( (x, z) \in A \subset \mathbb{R}^2 \) at time \( t \in \mathbb{R} \) via the following diffusion equation:

\[
\frac{\partial}{\partial t} T(x, t) = D_e(x) \Delta T(x, t) + \frac{P_{HVAC}(x, t) + e(x, t)}{\rho c_p}, \quad (5)
\]

where \( A \) is a closed domain of the atrium, \( D_e \) is the effective diffusion coefficient, \( P_{HVAC} \) is the heat input per unit time and unit volume from HVAC, \( e \) is the input from the other sources, \( \rho \) and \( c_p \) are the air density and the specific heat at constant pressure. \( P_{HVAC} \) is formulated as a function of outlet temperature \( T_{HVAC} \) (see Sec. 4.2).

4.2. Method for identifying \( D_e \)

This subsection provides the method for identifying \( D_e \).

The diffusion equation model (5) is discretized with the integer indices \( i \in \{1, 2, 3\} \) corresponding to TH1, TH2, and TH3, \( j \in \{1, \ldots, N\} \) to 3F–6F, and \( n = 0, \ldots, N - 1 \) to the time instant. Figure 5 shows the Staggered lattice [7] for discretization of the atrium. In this paper, the branches are denoted by \([i, j] \) and \([i, j_b] \) with the fractional indices \( i_b \in \{3/2, 5/2\} \) and \( j_b \in \{3/2, 5/2, 7/2\} \), and the effective coefficient \( D_e \) is defined on the branches. Note that the coefficient on \([j_b, j] \) is influenced by the heat input \( e(x, t) \) due to the in-room thermal dynamics.

Next, we derive Eq. (5) based on KMD. Let us represent the component of KM \( V_m^A \) on position \([i, j] \) with \( v_m[i, j] \) and that of \( V_m^H \) as \( u_m[i, j] \). It is shown in [3] that each oscillatory pattern \( \lambda^m \tilde{V}_m \) is determined by its initial condition and does not interact with any others. Thus, we assume that the oscillatory pattern is a solution of the discretized equation of (5). In addition, it is now possible to neglect the contribution of \( e(x, t) \) in Eq. (5) because \( e(x, t) \) determined by the outdoor temperature and solar radiation—varies on the day scale and does not contribute to the hour-scale thermal dynamics. Thus, by substituting \( \lambda^m \tilde{V}_m \) (or \( \lambda^m \tilde{u}_m \)) to \( T \) (or \( T_{HVAC} \)) in the discretized version of (5), the linear equation of \( v_m[i, j] \) and \( u_m[i, j] \) is derived as follows:

\[
s_m v_m[i, j] = \sum_{j=\pm 1/2} D_{b, j} \frac{v_m[i \pm 1, j] - v_m[i, j]}{L_n^2} + \sum_{j=\pm 1/2} D_{c, j} \frac{v_m[i, j \pm 1] - v_m[i, j]}{H^2} + \frac{U_{HVAC}}{V_0[i, j]} (u_m[i, j] - v_m[i, j]), \quad (6)
\]

where the two \( \pm \) signs of the first or second term of the right-hand side are in the same order, \( s_m := \ln[A_m]/\Delta t \) is the \( m \)-th eigen-angular frequency, and \( D \) with subscript of \( i, i_b, j, j_b \) is the effective coefficient. The parameter \( L_n \) is
the discretization step at the branches \([i_b, j]\) and independent on \(j\) because the temperature sensors are at the same locations on every floor. The constant \(H = 4.0 \text{ m}\) is the length between rooms, \(U_{\text{HVAC}} = 0.875 \text{ m}^3/\text{s}\) is outlet volume rate of HVAC units, and \(V_0[i, j]\) is the discretized volume of the node \([i, j]\) determined as shown in Figs. 1 and 2. The third term in the right-hand side of Eq. (6) is derived from the so-called bulk convection [1]. Moreover, we set a constraint condition for obtaining a physically-tractable solution of Eq. (6). Assuming that the air flow field is homogeneous because of no HVAC unit in the atrium, we introduce the following condition:

\[
D_{i_b,j} = \frac{D_{i+1,j} + D_{i-1,j} + D_{i,j+1} + D_{i,j-1}}{L_i^2 + L_j^2 + H^2}.
\]  

Eq. (7) minimizes the discretized form of \(\|\nabla D_{i,j}\|\) at the branch \([i, j]\), namely, the coefficient \(D_{i,j}\) does not drastically change in space. As a result, it is enough to solve Eqs. (6) and (7) for identifying the coefficient \(D_{i,j}\).

4.3. Result

Here, we calculate \(D_{i,j}\) by the method in Sec. 4.2 and verify the result with the characteristic numbers of fluid flow in the atrium and its architectural geometry. The dominant \(\tilde{V}_3\) was used for the calculation. Since the eigenperiod \(T_3 = 3.35 \text{ h}\) was small, it was reasonable in Eq. (6) to neglect the long-term dynamics. Also, we adopted the constraint condition (7) at the five nodes \([1,5,2,3,3/2,2,5/2,2,7/2,3,5/2]\) where the boundary conditions contributed less to the calculation than the other four nodes. Note that the coefficient matrix of linear equations (6) and (7) was not of full rank, and we hence used the corresponding pseudo-inverse matrix to determine \(D_{i,j}\).

Figure 6 shows the calculated coefficient \(D_{i,j}\) at the branch \([i, j]\). Note that the coefficient at \([i_0, j]\) reflects the in-room thermal dynamics and does not quantify the effective diffusion in the atrium. The calculated coefficient is in the order of \(10^{-2} \text{ m}^2/\text{s}\), which value is tractable because of the following reason. Now we set the target length-scale \(L\) of the heat transfer at \(H = 4.0 \text{ m}\), the in-room air velocity \(U\) at 0.8 m/s based on the outlet volume of HVAC units, and the molecular diffusion coefficient \(D\) at \(2.25 \times 10^{-3} \text{ m}^2/\text{s}\) [8]. Then the nominal value of the effective coefficient is given as \(\sqrt{DLU} = 8.5 \times 10^{-3} \text{ m}^3/\text{s}\) [5, 6] and close to the order of the calculated coefficient. This suggests that the effective heat diffusion by small-scale air movement is dominant in the target atrium.

Next, the space-dependency of the coefficient is discussed. The coefficient takes the large values at \([i, j_b] = [1, 3/2], [1, 7/2], [3, 5/2], [3, 7/2]\). This is partly because of the architectural geometry of the building. On the 3rd floor, a booth covers the center \([i, j_b] = [1, 3/2]\) of the atrium. Also, another office room is located on the 6th floor of the office room. Because these booth and room work as obstacle objects to the air flow, the coefficient is calculated as the large values at the nodes next to the objects.

5. Conclusion

This paper proposed a method for modeling of heat transfer dynamics in a building atrium via Koopman mode decomposition (KMD). First of all, KMD was applied to measured data on atrium temperature and outlet temperature of HVAC. Next, the heat transfer was modeled as a two-dimensional diffusion equation with an effective diffusion coefficient. Then, we proposed a method for identifying the coefficient by incorporating the spatio-temporal oscillatory pattern into the model. The coefficient calculated by the method shows that the small-scale air movement is dominant of the heat transfer in the atrium. Moreover, the space-dependency of diffusion is closely related to the architectural geometry of the atrium. Thus, the KMD-based method enables us to investigate the heat transfer driven by in-building HVAC systems.

References


On Smoothing Effects of Wind Power Via Koopman Mode Decomposition

Fredrik Raak†, Yoshihiko Susuki‡, Kazuhisa Tsuboki§, Masaya Kato§, and Takashi Hikihara†

†Department of Electrical Engineering, Kyoto University
Nishikyo, Kyoto 615–8510, Japan
‡Department of Electrical and Information Systems, Osaka Prefecture University
1–1 Gakuen-cho, Naka-ku, Sakai, Osaka 599–8531, Japan
§Hydrospheric-Atmospheric Research Center, Nagoya University
Chikusa, Nagoya 464–8601, Japan
Email: f-raak@dove.kuee.kyoto-u.ac.jp

Abstract—In this paper we apply a method known as Koopman Mode Decomposition (KMD) to data on simulated wind power outputs. We propose a new index which characterizes the smoothing of total output from multiple wind turbines or farms, through the spectral decomposition achieved via KMD. The index is experimentally exemplified via application to data where the maximum distance between measurement locations ranges from kilometers to hundreds of kilometers. Data on wind speeds from weather predictions obtained with CReSS (Cloud-Resolving Storm Simulator) are used together with a standard power curve to simulate outputs of wind turbines/farms. Results show that KMD reconstructs the wind farm output well by a set of modes, and that the smoothing resembles the conventional index based on power spectral densities for some cases. Furthermore, it is demonstrated with the KMD-based index that the smoothing on hour-scale for distributed wind farms in Japan exhibits similarity to the improved smoothing observed by distributing turbines over a larger area in a wind farm.

1. Introduction

In a system with large penetration of electric power from intermittent energy sources such as wind, it is important to distribute the generation over a large geographical area to smooth the total power generation. It thus becomes important to quantify the coherence of power generation at different locations when planning a suitable distribution of renewable power generation.

Here we will look at the smoothing of aggregated wind power, which has previously been addressed in e.g. [1–3]. In particular, [1] describes a way to estimate the Fourier spectrum of the total output of multiple turbines via the spectrum of a single one. Smoothing of output power from a Wind Farm (WF) through Power Spectral Densities (PSD) is also discussed in [2], where the authors analyzed data in Japan. In [3], a statistical analysis of correlation between wind power is conducted on very large scale, where it is shown that the correlation between aggregate wind power of large systems is similar to that of correlation between WFs.

In this paper, we look at the smoothing of wind power on different spatial and temporal scales by incorporating more and less detailed data on wind speeds from CReSS (Cloud-Resolving Storm Simulator) [4]. The detailed 0.5 Hz sampled data with a spatial resolution of 200 m are obtained in an area outside the coast of Aomori Prefecture in northern Honshu, Japan, while the more coarse 1 h−1 sampled data with 2 km resolution in Japan. To analyze the smoothing effects of wind power, we apply the so-called Koopman Mode Decomposition (KMD) (from its connection to the Koopman operator in dynamical systems theory) [5] to output powers of Wind Turbines (WT) or WFs, which transforms the time-series data into a finite number of modes evolving with single frequencies.

The contributions of this paper are mainly two-fold. First, a new and easily applicable index for the smoothing effects of wind power is proposed and compared with a conventional one, e.g. [2]. Second, it newly looks at the application of large-scale weather simulation data to analyze smoothing effects of wind power in Japan, which could be a viable method of analyzing potential sites for large-scale wind power development in the future.

2. Conventional Index of Wind Power Smoothing

This section revisits a conventional index of wind power smoothing [6]. We let \( S'(f) \) represent the PSD of wind speed at one WT in a WF, assuming that all WTs experience the same mean wind, and \( \gamma_{ij}(f) = \exp(-a d_{ij}/U) \) a function describing the coherency of wind speeds at the two measurement locations labeled by integers \( i \) and \( j \) that are separated by the distance \( d_{ij} \), where \( a \) is a decay constant, \( U \) mean wind speed, and \( f \) frequency. Note that the exponential function is an approximation of coherence for turbulence used in micrometeorology [6]. Thus, when considering a collective effect of winds over the WF with \( N \) turbines, the PSD of time-varying component of collective wind speed is described in [6] as follows:

\[
S_{WF}(f) = S'(f) \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_{ij}(f),
\]
where $\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_{ij}(f)$ is called the WF filter, which depends on the coherence functions and quantifies the degree of smoothing achieved at the WF.

Now considering wind power instead, the PSD of output power $P_W$ at a WF can be approximated in [2] as follows:

$$S_{P_W}(f) = S_P(f) \sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}(f) \cos(\phi_{ij}(f)), \tag{2}$$

similar to (1), where $S_P(f)$ represents the PSD of output power at a typical WT, $\phi_{ij}(f)$ a function determining the phases of power outputs, and $\Gamma_{ij}$ the coherence function. The WF filter here becomes $G_{WF}(f) := S_{P_W}(f)/S_P(f) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} \Gamma_{ij}(f) \cos(\phi_{ij}(f))$. The coherence function $\Gamma_{ij}(f)$ is in [2] as follows:

$$\Gamma_{ij}(f) = \frac{|S_{P_i}(f)|}{\sqrt{S_{P_i}(f)S_{P_j}(f)}}, \tag{3}$$

which quantifies the magnitude of overlapping frequency content of two signals, where $S_{P_i}$ is the PSD of electric power at WT $i$, and $S_{P_j}$ the cross spectral density between $i$ and $j$. It is shown in [2] that the exponential approximation of (3) like above offers a good agreement with experimentally-obtained data.

Suppose that output powers $P_i$ are measured at every WT $i$, and the total power corresponds to $P_{tot} = \frac{1}{m} \sum_{i=1}^{m} P_i$ in per unit (p.u.). The following smoothing index with respect to a typical WT $i$ is then derived:

$$s_i(f) := \sqrt{S_{P_i}(f)/(S_P(f))}, \tag{4}$$

which is the gain of $G_{WF}(f)$ in accordance with [2], as a comparison to (9) proposed later for KMD. The functions $S_{P_i}$ and $S_P$ are the PSDs of $P_{tot}$ and $P_i$. Since a "typical WT" could be a difficult task to determine, in particular when outputs and locations vary significantly, we consider calculating the mean smoothing according to

$$s(f) := t\text{mean}(s_i(f)), \tag{5}$$

where $t\text{mean}$ represents the 25% truncated mean over all WTs $i = 1, \ldots, m$, to remove outliers.

3. New Smoothing Index via Koopman Modes

In this section, we introduce a new index to characterize the smoothing effects of wind power based on the so-called Koopman Mode Decomposition (KMD). We refer to [5, 7–9] for detailed theoretical background of KMD.

Now, we consider $N + 1$ vector-valued snapshots of wind power measurements collected at $m$ locations: $\{P_0, \ldots, P_N\}$, $P_k \in \mathbb{R}^m$. The sampled time-series are then decomposed into a finite sum via KMD:

$$P_k = \sum_{i=1}^{N} \tilde{\lambda}_i^k \tilde{v}_i, \quad k = 0, \ldots, N-1, \tag{6}$$

$$P_N = \sum_{i=1}^{N} \tilde{\lambda}_i^N \tilde{v}_i + r,$$

computed via an Arnoldi-type algorithm [5], giving $N$ pairs of so-called Ritz-values $\tilde{\lambda}_i \in \mathbb{C}$ and vectors $\tilde{v}_i \in \mathbb{C}^m$. The vector $r$ is the residual component in KMD; if assumed to be zero, (6) becomes

$$P_k = \sum_{i=1}^{N} \tilde{\lambda}_i^k \tilde{v}_i, \quad k = 0, \ldots, N. \tag{7}$$

Frequencies are calculated as $f_i = \text{Im}(\ln(\tilde{\lambda}_i))/(2\pi T_s)$, where $T_s$ is the sampling period of data. The vector $\tilde{v}_i$ is here called the Koopman Mode (KM) and contains the magnitudes and phases of power fluctuations at the measurement locations for the frequency $f_i$. To identify lightly damped or undamped oscillations with large magnitude, all $N$ KMs are sorted by $(\tilde{\lambda}_i)^N \|\tilde{v}_i\|$, and higher ranked ones are called dominant KMs. Here, KMD will be applied to wind powers at $m$ locations, representing hypothetical WTs or WFs. The total power $P_{tot,k}$ can be expressed using (7) as

$$P_{tot,k} = \sum_{i=1}^{N} \tilde{\lambda}_i^k \sum_{j=1}^{m} [\tilde{v}_i]_j = \sum_{i=1}^{N} \tilde{\lambda}_i^k \tilde{v}_i, \tag{8}$$

where $\tilde{v}_i \in \mathbb{C}$ is the scalar KM of the total output. That is, a spectral decomposition of the total output power is achieved through the decomposition of individual outputs. Now let $\tilde{A}_i = |\tilde{v}_i|$ be the amplitude factor the $i$-th KM oscillation for the total output, and analogously $A_{ij} = |\tilde{v}_{ij}|$ the factor for the same oscillation for a WT or WF $j$, and $A_i = [A_{i1}, \ldots, A_{im}]^\top$ (T denotes vector transpose.) Then we define

$$s_i := \tilde{A}_i/(m \cdot t\text{mean}(A_i)), \tag{9}$$

as a index of smoothing with respect to frequency $f_i$, where $t\text{mean}$ is same as in (5), taken over all $m$ turbines, and $m$ is included in (9) to scale down the total output to p.u. If the amplitude of an oscillation is smaller in the sum than for individual WTs or WFs, then (9) becomes smaller than one for that particular $f_i$. 

Figure 1: Placement of wind turbines (red dots) in a hypothetical wind farm with 600 m distance between turbines. The colored background and arrows indicate the speed and direction of wind. The black curve depicts the coastline of Aomori Prefecture.
Figure 2: (a) Turbine outputs for 600 m (distance between consecutive turbines) case; (b) wind farm outputs and approximations by dominant KMs; smoothing results via (c) KMD and (d) PSDs.

4. Demonstration

Now, two examples are used to evaluate the index proposed in this paper. The first one incorporates 0.5 Hz sampled wind prediction data from the CReSS weather model [4], averaged over a grid with 200 m spatial resolution. The data are used as input to WTs in a hypothetical WF outside the coast of Aomori Prefecture in Japan [10]. A single array of 15 turbines is considered, and three cases are considered where the distances between consecutive turbines are chosen as 400, 600, and 800 m; see red dots in Fig. 1 which depicts the 600 m case. The “lowest” WT position in the figure is common for all cases. Wind speeds are converted into power outputs with a standard power curve. The KMD and proposed index are applied to sampled powers for 16 minutes, where wind speeds are dramatically increasing due to an incoming winter storm, and turbine outputs are shown in Fig. 2 (a) for the 600 m case.

Fig. 2 (b) shows how well the subsets of dominant KMs approximate the wind farm outputs using (8). Smoothing results for all cases calculated with (9) are presented in Fig. 2 (c). The results indicate considerable smoothing that increases with higher frequency. In particular, for frequencies higher than about $4 \times 10^{-3}$ Hz, i.e. periods less than 4 min, $s_i$ becomes small, indicating the wind farm smooth-
ing. As a comparison, the smoothing index \( s(f) \) calculated via PSDs is given in Fig. 2 (d). The results on PSD-based smoothing initially looks similar and then increases towards and above unity for all cases. This would imply that a certain fluctuation in the output of a WT is magnified by the wind farm, which is not realistic. If all WTs e.g. oscillate 1 p.u. in unison for a particular frequency, the wind farm output would also oscillate with 1 p.u. (with respect to its maximum capacity.) The issue here may be that the different WTs have significantly different spectrum, thus it becomes difficult to average the results or estimate the smoothing based on a single turbine output.

On the other hand, the KMD-based smoothing result indicates that the smoothing significantly improves for the 800 or the 600 m distance case between WTs compared with the 400 m case, in difference to the PSD-based index which is inconclusive on this point. We see that this seems to agree with the observed WF outputs in Fig. 2 (b) since e.g. the WF output in the 800 m case fluctuates less than for the 400 m case.

Now we look at more coarse CReSS data of whole Japan (2 km spatial resolution and \( T_r = 1 \) h). We consider 41 hypothetical wind farms (or locations) whose outputs are calculated with the same power curve used previously. The locations vary from a WF distribution concentrated to the northern part of Honshu (Case 1) to sparsely distributed WFs all along Japan’s coastline (Case 5). The two most extreme cases (1 and 5) are shown in Fig. 3 and highlighted by red and black crosses. The total outputs of all wind farms for the all cases are given in Fig. 4 (a), together with the reconstructed outputs by dominant KMs, and smoothing results are given in Figs. 4 (b) and (c).

In this case, the KMD-based smoothing index shows more resemblance to the results achieved via PSD-based smoothing, although indicating slightly more smoothing. The cause of more resemblance here might be that the spectrum at different locations are more homogeneous, thus the smoothing based on the mean or individual PSDs become similar. According to the KMD-based smoothing, the improved smoothing on this large scale resembles the improvement achieved in the previous example for a WF (compare Fig. 2 (c) and Fig. 4 (c)), by distributing WTs over a larger area, however more data should be considered in the future to validate this. This scale invariance is also supported by the results of e.g. [3, 11]. In particular [11] shows that the reduced variability of summed power outputs of WFs is comparable to that of individual WFs.

5. Conclusions

A new smoothing index of aggregated wind power was proposed based on measurements of power at each wind farm or turbine, via the so-called Koopman Mode Decomposition. The index was applied to wind turbine outputs in a hypothetical wind farm in an area attractive for wind power in Japan, as well as to coarse data of whole Japan. According to the proposed index, the improved smoothing on hour-scale for distributed wind farms in Japan becomes similar to the improved smoothing by distributing turbines over a larger area in a wind farm. However, an investigation including more data should be conducted to validate the correctness of this, which is one of the future works. The KMD-based index and conventional power spectral density-based one become similar for one of two cases here. We speculate that the agreement is dependent on the degree of homogeneity of the spectrum of turbines and farms, which could make it difficult to average the smoothing result or select an appropriate comparison to the aggregated wind power output.

References

Optimal Parameter Selection for Kernel Dynamic Mode Decomposition

Wataru Kurebayashi†, Sho Shirasaka‡ and Hiroya Nakao††

†Dept. of Information Engineering, Aomori University
2–3–1 Kohbata, Aomori, Aomori 030-0943, Japan
‡Graduate School of Information Science and Engineering, Tokyo Institute of Technology
||School of Engineering, Tokyo Institute of Technology
2–12–1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
Email: wkurebayashi@gmail.com

Abstract—The dynamic mode decomposition is a recently proposed algorithm for the mode decomposition of mixed-mode time series data based on the dynamical systems theory. The kernelization of this algorithm improves the estimation accuracy of dynamic modes, and therefore, facilitates wide-range applications of this method, while the optimal hyper-parameter selection for kernel dynamic mode decomposition still remains an open question. Here, we propose a formula for the hyper-parameter selection, and demonstrate the validity of our selection method.

However, the selection of optimal hyper-parameters of the extended and kernel DMDs still remains an open question. Since the selection of hyper-parameters deeply affects the estimation accuracy of the eigenvalues and eigenfunctions of the Koopman operator, it is important to establish the selection method for hyper-parameters. In this paper, we propose a selection method [4] for hyper-parameters of the kernel DMD, and demonstrate the validity of this method through numerical experiments.

1. Introduction

Dynamic mode decomposition (DMD) [1] is a recently proposed mode decomposition method for mixed-mode time series data. The advantage of this method is that it can directly take into account the latent dynamics of time series. In the field of fluid mechanics, for example, a conventional method called the proper orthogonal decomposition has been widely used, which performs the principal component analysis to reduce the dimensionality of time series, and does not take into account the latent dynamics. Since the DMD can capture the latent dynamics, it is useful for modeling the latent dynamics in terms of dynamical systems theory.

Although the computation of dynamic modes in the DMD is no more than a linear transformation, i.e., the DMD neglects the nonlinearity of time series, the theoretical basis of the DMD based on the Koopman operator can deal with nonlinear latent dynamics underlying time series, so the nonlinear extension of the DMD is expected to significantly enhance the potential of this method. Then, the extended DMD [2] and the kernel DMD [3] have been proposed. The extended DMD [2] is a modified version of the DMD that can approximate the Koopman operator more precisely by nonlinear regression. In addition, the kernel DMD [3] improves the computational efficiency of the extended DMD by what we call “kernel trick”. Although the original DMD needs significantly high-dimensional time series data, e.g., simulation data in fluid mechanics, the extended and kernel DMDs can be applied to a wider class of time series. Therefore, these extensions are expected to significantly widen the applicability of the DMD.

2. Kernel dynamic mode decomposition

To begin with, we introduce a linear operator called the Koopman operator that gives a theoretical basis to the DMD. The DMD is an algorithm that computes the eigenvalues and eigenfunctions of the Koopman operator, which are used for the mode decomposition of time series.

We consider a discrete-time stationary dynamical system, described by

\[ x_{t+1} = F(x_t), \]  

where \( x_t \in \Omega \) is a vector of state variables at time \( t \) and \( F(x) \) is a map that represents the dynamics of this system. Here, let \( f(x) (x \in \Omega) \) be an element of an appropriate functional space \( F \). The Koopman operator \( \mathcal{K} \) corresponding to the system of Eq. (1) is defined as follows:

\[ \mathcal{K}f(x) = f(F(x)), \]

that is, the Koopman operator \( \mathcal{K} \) maps any function \( f(x) \in F \) to a composite function \( f(F(x)) \). This operator can be interpreted as a time-shift operator that acts on an observable \( f(x) \), because the following relation holds:

\[ \mathcal{K}f(x)|_{x=x_t} = f(x_{t+1}). \]

The definition of the Koopman operator can be naturally extended to stochastic dynamical systems. Instead of Eq. (1), we here consider a stationary Markov process:

\[ x_{t+1} \sim p(x_{t+1}|x_t) \]
In this case, the Koopman operator is defined as a conditional expectation as follows:

\[ Kf(x) = \int_\Omega f(x')p(x'|x)dx' \tag{5} \]

Since the Koopman operator is a linear operator, we can consider the spectral decomposition of this operator. In the DMD, the Koopman operator is approximated only by discrete eigenpairs called point spectrum. Each discrete eigenvalue \( \lambda_i \) (i = 1, 2, 3, ...) of the Koopman operator has a corresponding eigenfunction \( \xi_i(x) \) such that

\[ K\xi_i(x) = \lambda_i \xi_i(x). \tag{6} \]

By using the eigenfunctions \( \{\xi_i(x)\} \), the time series of state variables \( x_t \) can be decomposed as follows:

\[ x_t = \sum_{i=1}^{\infty} c_i \xi_i(x_t), \tag{7} \]

where \( \{c_i \in \Omega\} \) is appropriate coefficients such that \( x = \sum_i c_i \xi_i(x) \). In addition, if the dynamics of \( x_t \) is deterministic, the time series \( x_t \) can be written as a sum of sinusoidal waveforms as follows:

\[ x_t = \sum_{i=1}^{\infty} c_i \lambda_i^t \xi_i(x_0). \tag{8} \]

Note that we neglected decomposition components in Eq. (8) under the assumption of stationarity.

In the kernel DMD, the eigenvalues \( \{\lambda_i\} \) and eigenfunctions \( \{\xi_i(x)\} \) of the Koopman operator can be computed from time series data of state variables \( \{x_1, x_2, \ldots, x_n\} \) by the following procedure. This algorithm has two hyper-parameters, the kernel parameter \( \theta_k \) and the regularization parameter \( \theta_r \). The kernel parameter represents the width of a kernel function. For example, we can use the Gaussian kernel with bandwidth parameter \( \theta_k \), i.e., \( k(x, x') = \exp(-\frac{1}{2\theta_k} \|x-x'\|^2) \).

1. Computation of \((n-1) \times (n-1)\) matrices \( K, L \) whose \((i, j)\)-th element \( K_{ij}, L_{ij} \) is given by

\[ K_{ij} = k(x_i, x_j), L_{ij} = k(x_i, x_{j+1}). \tag{9} \]

2. Singular value decomposition of \( K' \):

\[ K = U\Sigma V^T, \tag{10} \]

where \( U = [u_1, \ldots, u_{n-1}], V = [v_1, \ldots, v_{n-1}] \) are orthogonal matrices, and \( \Sigma \) is a diagonal matrix whose diagonal elements are \( \sigma_1 \geq \ldots \geq \sigma_{n-1} \).

3. Let \( n' \) be the number of singular values that covers \( 100 - \theta_r \) percent of the square sum of all singular values of \( K \). We approximate \( K \) by the \( n' \) largest singular values as

\[ K \approx \tilde{U} \tilde{\Sigma} \tilde{V}^T \tag{11} \]

where \( \tilde{U} = [u_1, \ldots, u_{n'}], \tilde{V} = [v_1, \ldots, v_{n'}] \), and \( \tilde{\Sigma} \) is a diagonal matrix whose diagonal elements are \( \sigma_1 \geq \ldots \geq \sigma_{n'} \).

4. Computation of \( \tilde{A} (= \tilde{U}^T LK^{-1} \tilde{U}) \):

\[ \tilde{A} = \tilde{U}^T L \tilde{V} \tilde{\Sigma}^{-1}. \tag{12} \]

5. Eigendecomposition of \( \tilde{A}^T \). Let \( \{\lambda_i\}, \{q_i\} (i = 1, \ldots, n') \) be the eigenvalues and eigenvectors of \( \tilde{A}^T \).

6. Computation of the eigenfunctions of the Koopman operator, \( \{\xi_i(x)\} \), as follows:

\[ \xi_i(x) = q_i^T \tilde{U}^T \phi(x), \tag{13} \]

where \( \phi(x) = [k(x, x_1), \ldots, k(x, x_{n-1})]^T \).

Note that we introduced the regularization with hyperparameter \( \theta_r \) (steps 2–4), which is not included in Williams et al [3].

3. Criterion for parameter selection

In this section, we derive a criterion for selecting two hyper-parameters of the kernel DMD, the kernel parameter \( \theta_k \) and the regularization parameter \( \theta_r \). We evaluate the error between the exact and estimated Koopman operators, and select the two hyper-parameters that minimize the estimation error.

In order to evaluate the estimation error, we introduce the mean square error between the exact and estimated Koopman operators, \( \mathcal{K} \) and \( \tilde{\mathcal{K}} \), as follows:

\[ \text{MSE} = \int_\Omega \int_\Omega \{p(x|x') - \tilde{p}(x|x')\}^2 q(x')dx'dx'. \tag{14} \]

where \( \tilde{p}(x|x') \) is the estimated conditional probability assumed implicitly in the algorithm of the kernel DMD, and \( q(x) \) is the stationary probability density of the state \( x \). In Eq. (14), we calculate a square error \( \left(\int_\Omega \{p(x|x') - \tilde{p}(x|x')\}^2 q(x')dx'\right) \) and its expectation value with respect to \( x' \), so we call Eq. (14) the mean square error. The advantage of this criterion is that it is plausible to approximately compute it only from data. For example, the Hilbert-Schmidt norm of \( \mathcal{K} - \tilde{\mathcal{K}} \) is a possible alternative of Eq. (14), but it is difficult to compute it only from data.

Equation (14) cannot be evaluated only by time series data, because it includes the unknown exact probability density \( p(x|x') \). Then, from Eq. (14), we derive a cost function that can be evaluated only by time series data. We can rewrite Eq. (14) as follows:

\[ \text{MSE} = \int_\Omega \int_\Omega \{p(x|x') - \tilde{p}(x|x')\}^2 q(x')dx'dx' 
- 2 \int_\Omega \int_\Omega \tilde{p}(x|x')p(x|x')q(x')dx'dx' 
+ \int_\Omega \int_\Omega \{\tilde{p}(x|x')\}^2 q(x')dx'dx'. \tag{15} \]
Since the first term of the right hand side of Eq. (15) does not depend on the hyper-parameters, we define a cost function $C$ as follows:

$$C = \text{MSE} - \int_{\Omega} \int_{\Omega} \{p(x|x')\}^2 q(x') dx' dx'$$

$$= -2 \int_{\Omega} \int_{\Omega} p(x|x') \hat{p}(x|x') q(x') dx' dx'$$

$$+ \int_{\Omega} \int_{\Omega} \{\hat{p}(x|x')\}^2 q(x') dx' dx'. \quad (16)$$

The cost function $C$ can be approximately evaluated by time series data $\{\tilde{x}_1, \ldots, \tilde{x}_n\}$ that is not used for the estimation of the Koopman operator as follows:

$$\hat{C} = -\frac{2}{n-1} \sum_{i=1}^{n-1} \hat{p}(\tilde{x}_{i+1}|\tilde{x}_i) + \frac{1}{n} \sum_{i} \{\hat{p}(x|\tilde{x}_i)\}^2 dx. \quad (17)$$

Here, the estimated conditional probability $\hat{p}(x|x')$ assumed implicitly in the kernel DMD can be derived as follows (see Ref. [4] for derivation):

$$\hat{p}(x|x') = \phi(x)^\top \hat{U} (\hat{U}^\top \hat{G} \hat{U})^{-1} \hat{A} \hat{U}^\top \phi(x'), \quad (18)$$

where $G$ is an $(n-1) \times (n-1)$ matrix whose $(i,j)$-th element is given by $\int_{\Omega} k(x, x_i) k(x, x_j) dx$. By using Eq. (18), the cost function $\hat{C}$ can be written as

$$\hat{C} = -\frac{2}{n-1} \sum_{i=1}^{n-1} \phi(\tilde{x}_{i+1}|\tilde{x}_i) \hat{U} (\hat{U}^\top \hat{G} \hat{U})^{-1} \hat{A} \hat{U}^\top \phi(\tilde{x}_i)$$

$$+ \frac{1}{n} \sum_{i} \phi(\tilde{x}_i)^\top \hat{U} \hat{A} \hat{U}^\top (\hat{U}^\top \hat{G} \hat{U})^{-1} \hat{A} \hat{U}^\top \phi(\tilde{x}_i). \quad (19)$$

Thus, we can select the hyper-parameters, $\theta_k$ and $\theta_r$, by evaluating the cost function $\hat{C}$ for each parameter set and selecting the parameter set $\theta_k, \theta_r$ that minimizes $\hat{C}$. The evaluation of $\hat{C}$ can be performed by the $m$-fold cross validation of Eq. (19) as follows:

1. Divide the time series $\{x_i\}$ into $m$ subsets, denoted by $X_i$ ($i = 1, 2, \ldots, m$).
2. For $i = 1, 2, \ldots, m$:
   (a) Apply the kernel DMD to all the time series except $X_i$ and compute $\hat{A}, \hat{U}$ and $\hat{G}$.
   (b) Evaluate the cost function $\hat{C}$ of Eq. (19) by $\hat{A}, \hat{U}, \hat{G}$ and $X_i$.
3. Average the values of $\hat{C}$ evaluated for $X_1, \ldots, X_m$.
4. Numerical experiments

In order to confirm the validity of our method, we perform a numerical experiment, in which we apply our method to time series data generated by a numerical simulation. As an example, we use two-dimensional time series data \( x_t = [x_t^{(1)}, x_t^{(2)}]^T \) depicted in Fig. 1(a,b). This time series is generated from a simple dynamical system described by \( x_t^{(1)} = \exp(\rho_1 + \sin 2\varphi_t) \cos(\varphi_{t+1} + \rho_1 + \sin 2\varphi_t), \quad x_t^{(2)} = \exp(\rho_2 + \sin 2\varphi_t) \sin(\varphi_{t+1} + \rho_2 + \sin 2\varphi_t), \)

\( \varphi_{t+1} = \varphi_t + \pi/15 + 0.1\eta_{t}^{(1)}, \quad \) and

\( \rho_{t+1} = 0.9\rho_t + 0.1\eta_{t}^{(2)}, \)

where \( \eta_{t}^{(1)}, \eta_{t}^{(2)} \) are independent Gaussian white noise of unit intensity. In this case, we can analytically derive the Koopman eigenfunctions as shown in Figs. 2–4(a).

We show the evaluation of the cost function \( \hat{C} \) in Fig. 1(c), which is evaluated by 5-fold cross validation. The parameter sets A, B and C shown in Fig. 1(c) represent optimal, narrower and wider kernel parameters, which will be used as examples in the following discussion.

In Figs. 2–4, we show the eigenfunctions computed by the kernel DMD, which correspond to dynamical variables such as a phase or amplitude of oscillatory modes. In each figure, we compare (a) the exact eigenfunction with (b–d) the estimated eigenfunctions for the parameter sets (b) A, (c) B and (d) C. In each of Figs. 2–4, we see that the results for the parameter set A nicely fits to the exact eigenfunction, while the result for the parameter set B (narrower kernel) is noisy and overfitted, and the result for the parameter set C (wider kernel) is underfitted. Note that the region with no data point is shown in white, because the kernel DMD cannot estimate the eigenfunction in this region. From these results, we see that our method can find the optimal hyper-parameter for this time series data.

5. Summary

We proposed a method for selecting two hyper-parameters, \( \theta_k, \theta_r \) [4], which deeply affect the estimation accuracy of Koopman eigenfunctions. In a numerical experiment, we demonstrated the validity of our method. By using the optimal hyper-parameters, we could estimate eigenfunctions of the Koopman operator in high precision. This result implies the importance of the optimal selection of hyper-parameters of the kernel DMD.

Acknowledgment

W.K. is financially supported by JSPS KAKENHI JP16K16125 (Grant-in-Aid for Young Scientists (B)). S.S. is financially supported by JSPS KAKENHI JP15J12045. H.N. is financially supported by JSPS KAKENHI JP16K13847, JP16H01538.

References

Revisiting delay embedding: dynamical reconstruction based on Sturm-Liouville theory

Naoto Nakano†‡

†JST PRESTO, 4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan
‡Faculty of Science, Hokkaido University, Kita 10 Nishi 8, Kita-ku, Sapporo 060-0810, Japan
Email: n_nakano@math.sci.hokudai.ac.jp

Abstract—Delay embedding is well-known for non-linear time-series analysis, and it is used in several research fields. Takens theorem ensures validity of the delay embedding analysis: embedded data preserves topological properties, which the original dynamics possesses, if one embeds into some phase space with sufficiently high dimension. This means that, for example, an attractor can be reconstructed by the delay coordinate system topologically. However, configuration of embedded data may easily vary with the delay width and the delay dimension, namely, “the way of embedding”. In a practical sense, this sensitivity may cause degradation of reliability of the method, therefore it is natural to require robustness of the result obtained by the embedding method in certain sense.

In this study, we investigate the mathematical structure of the framework of delay-embedding analysis to provide Ansatz to choose the appropriate way of embedding, in order to utilise for time-series prediction. In short, mathematical theories of the Hilbert–Schmidt integral operator and the corresponding Sturm–Liouville eigenvalue problem underlie the framework. Using those mathematical theories, one can derive error estimates of mode decomposition obtained by the present method and a time evolution equation represented by the mode amplitude functions constructed exclusively by given time-series. Moreover, projecting datasets into a subspace spanned by the leading modes, we can detect the attractor and analyse the corresponding dynamics. In this talk, we will show some results for some numerical and experimental datasets to validate the present method.

In fact, this mathematical justification relies on the $L^2$ analysis and the modes of the decomposition corresponds to intrinsic modes of the autocorrelation function, namely, the intrinsic frequency modes. Hence, this methodology is expected to have some relevance to the Koopman mode decomposition, which is used to extract characteristic frequency of the signal.
A Schrödinger-type Formalism and Observable Wavefunctions in Dynamical Systems

Igor Mezić†

†Departments of Mechanical Engineering and Mathematics, University of California, Santa Barbara, CA 93106-5070, USA
Email: mezic@engineering.ucsb.edu

Abstract—There is recent interest in the use of Koopman (composition) operator theory for a wide range of problems in dynamical systems. In its dual, Perron-Frobenius theory, the use of invariant measures for understanding of statistical properties of dynamical systems is routine. A much less used concept is that of eigenmeasures [10]. We extend the theory related to eigenmeasures to introduce the notion of wavefunctions into dynamical systems theory. A wavefunction can be thought of as the density of a complex measure on the state space. It satisfies the common Perron-Frobenius equation. Using this, we derive a Shrödinger-type formalism for complex measure propagation on embeddings of dynamical system dynamics into the output space of an observable propagated by the Koopman operator. The resulting wavefunction is named an observable wavefunction (OW).

1. Introduction

Driven by success in operator-based framework in quantum theory, Bernard Koopman proposed in his 1931 paper [3] to treat classical mechanics in a similar way, using the spectral properties of an operator associated with the dynamical system evolution. Koopman extended this study in a joint work with von Neumann in [2]. Those works, restricted to Hamiltonian dynamical systems, did not attract much attention originally, as evidenced by the fact that between 1931 and 1963, the Koopman paper [3] was cited 25 times, according to Google Scholar. This can be attributed largely to the major success of the geometric picture of dynamical systems theory in its state-space realization advocated by Poincaré. In fact, with Lorenz’s discovery of a strange attractor in 1963, the dynamical systems community turned to studying dissipative systems and much progress has been made since. Within the current research in dynamical systems, some of the crucial roadblocks are associated with high-dimensionality of the problems and necessity of understanding behavior globally (away from the attractors) in the state space. However, the weaknesses of the geometric approach are related exactly to its locality (it often relies on perturbative expansions around a known geometrical object) and low-dimensionality (it is hard to make progress in higher dimensional systems using geometry tools).

Out of today’s 400+ citations of Koopman’s original work, [3], 75% come from the last 20 years. Thus, it was only in the 1990’s that potential for wider applications of the operator-theoretic approach has been realized [4, 7]. In this century the trend of applications of this approach has continued, as summarized in [1]. This is partially due to the fact that strong connections have been made between the spectral properties the Koopman operator for dissipative systems and the geometry of the state space. In fact, the hallmark of the work on the operator-theoretic approach in the last two decades is the linkage between geometrical properties of dynamical systems - whose study has been advocated and strongly developed by Poincaré and followers - with the geometrical properties of the level sets of Koopman eigenfunctions [7, 5, 6]. The operator-theoretic approach has been shown capable of detecting object of key importance in geometric study, such as invariant sets, but doing so globally, as opposed to locally as in the geometric approach. It also provides an opportunity for study of high-dimensional evolution equations in terms of dynamical systems concepts [8, 11] via a spectral decomposition, and links with associated numerical methods for such evolution equations [12].

In this paper we consider the propagation of observables under the Koopman operator and derive an equation for wavefunction evolution for such a propagation. We first...
Define the notion of a wavefunction on the state space. Then we define a complex observable on the state space and consider the evolution of a complex measure associated with such an observable. The result is a Schrödinger-type formalism that couples the Koopman operator and the Peron-Frobenius operator state-space formalisms and extends them to embedding space of observable outputs. We pursue this in 1−D here, and will present the n-dimensional theory in a forthcoming paper.

2. Wavefunctions for Observable Evolution

Let \( M = \mathbb{R} \) and \((x,t) \in \mathbb{R} = \mathbb{R} \times \mathbb{R} \). Let \( v \) be a smooth vector field on \( \mathbb{R} \). The wavefunction \( \rho \) (we will call it the true wavefunction or TW) satisfies

\[
\frac{\partial \rho}{\partial t} + \frac{\partial \rho v}{\partial x} = 0, \tag{1}
\]

Let the observable \( f : \mathbb{R} \to \mathbb{C} \) be defined by

\[
f = e^{-iY}, \tag{2}
\]

where \( Y(x,t) \) is smooth (at least in \( C^2 \)). This implies that the observable wavefunction (OW) \( \psi \) defined by

\[
\psi = \frac{\rho}{iY e^{iY}}, \tag{3}
\]

is the density of a complex measure of the observable \( f = e^{-iY} \) corresponding with the TW \( \rho \), since

\[
\frac{df}{dx} = \frac{\partial Y}{\partial x} e^{iY}. \tag{4}
\]

We proceed to derive an equation of evolution for \( \psi \). We will denote partial derivatives with respect to \( t \) by \( \cdot_t \) and partial derivatives with respect to \( x \) by \( \cdot_x \).

We have

\[
\psi_t = \frac{\rho_t}{iY_x e^{iY}} + \rho \left( \frac{1}{iY e^{iY}} \right)_t
\]

\[
= -\frac{v \rho_x}{iY_x e^{iY}} - \frac{v \rho}{iY e^{iY}} + \rho \left( \frac{1}{iY e^{iY}} \right)_t. \tag{5}
\]

From (3) we have

\[
\rho_x = iY_x \psi e^{iY} + iY_x \psi e^{iY} - \psi(Y_x)^2 e^{iY}, \tag{6}
\]

and, as a consequence,

\[
-\frac{v \rho_x}{iY_x e^{iY}} = -\frac{v}{Y_x} \psi \phi Y_x - \frac{v}{Y_x} (\psi(Y_x)^2). \tag{7}
\]

We also have

\[
\left( \frac{1}{iY_x e^{iY}} \right)_t = -\left( \frac{1}{iY_x e^{iY}} \right) \left( Y_{xt} e^{iY} - Y_x Y_t e^{iY} \right) = -\left( \frac{Y_{yt}}{i(Y_x)^2 e^{iY}} + \frac{Y_t}{Y_x e^{iY}} \right), \tag{8}
\]

and thus

\[
\rho \left( \frac{1}{iY_x e^{iY}} \right)_t = -iY_x \psi e^{iY} \left( \frac{Y_{yt}}{i(Y_x)^2 e^{iY}} + \frac{Y_t}{Y_x e^{iY}} \right) = -\psi \left( \frac{Y_{yt}}{Y_x} + iY_t \right). \tag{9}
\]

Collecting all of these, from (5) we get

\[
\psi_t = -v \psi_x - \psi_v - \frac{v}{Y_x} Y_x \psi - \frac{Y_{yt}}{Y_x} \psi - i \left( \frac{Y_t}{Y_x} \right) \psi. \tag{10}
\]

This is the equation that governs the observable wavefunction evolution. If the observable is real, with \( Y = iK \), we get

\[
\psi_t = -v \psi_x - \psi_v - \frac{v}{K_x} K_{xx} \psi - \frac{K_{yt}}{K_x} \psi - (K_t + \frac{v}{K_x} (K_x)^2) \psi = -v \psi_x - \psi_v - (K_t - \frac{v}{K_x} (K_{xx} - K_x^2 - K_x^2) \psi).
\]

3. Conclusions

We have derived a wavefunction formalism for continuous-time dynamical systems in 1D. The theory developed here leads to Schrödinger-type equations for evolution of constant speed on a 1-dimensional Riemannian manifolds [9]. It also admits generalization to higher dimensions. The formalism provides a coupling between Koopman operator theory - evolving observables - and Schrödinger operator theory - evolving densities - for embeddings of dynamical systems.

Acknowledgments

The author would like to thank NOLTA2016 organizing committee members for their fruitful suggestions and comments. This research was partially supported by the ARO Grant W911NF-14-1-0359.
References


Lyapunov Bundle on 0-Dimensional Section Torus and Bifurcation of Quasi-Periodic Solution

Kyohei Kamiyama‡ and Motomasu Komuro‡ and Tetsuro Endo§ and Kazuyuki Aihara‡

‡Institute of Industrial Science, The University of Tokyo
4–6–1 Komaba, Meguro-ku, Tokyo, 153–8505, Japan
‡Center for Fundamental Education, Teikyo University of Science,
2525 Yatsusawa, Uenohara-shi, Yamanashi, 409–0193, Japan
§Department of Electronics and Bioinformatics, School of Science & Technology, Meiji University
1–1–1 Higashi-Mita, Tama-ku, Kawasaki-shi, Kanagawa, 214–8571 Japan
Email: kamiyama@sat.t.u-tokyo.ac.jp, komuro@ntu.ac.jp, endoh@meiji.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—Quasi-periodic solutions are ordinary phenomena in coupled oscillatory systems, such as electrical oscillators, metronomes, fireflies, frogs, and also quantum computer. The quasi-periodic solution attracts many researchers in recent years. However, its bifurcation analysis was difficult; this is because it has no period. Recently, we have developed a bifurcation analysis tool called Lyapunov bundle which is a set of Lyapunov vectors on a torus. The mechanism of local bifurcation of quasi-periodic solution is explained by the topology of dominant Lyapunov bundle. The topology of Lyapunov bundle on high-dimensional torus is difficult to analyze as it is. Therefore, we took a slice called one-dimensional section torus (ST0) on the original torus and developed a bifurcation analysis tool for higher-dimensional torus named Lyapunov bundle (abbr. LB) [1], [2]. The LB is a generalization of eigenvectors to quasi-periodic solutions and a set of tangent vectors, called Lyapunov vectors (abbr. LV), on every points in a torus. Local bifurcation type is clarified by the topology of LB around a local bifurcation point.

For now, the classification of the topology of LB has been done by visual, not programmatically. Hence, the classification becomes difficult, if dimension is higher or shape of torus is complex. Therefore, we decrease the dimension by one by taking Poincare section for continuous-discrete time system or we decrease the dimension by one by taking the intersection space with \( n – 1 \)-dimensional hyperplane for discrete-time dynamical system. However, even if we use this decreasing method, we can classify up to 2-dimensional torus at the most. In order to this problem, we take 0-dimensional section from the original torus and its LB. Furthermore, we develop a classification method for LB on 0-dimensional section torus (ST0).

1. Introduction

In the real world, there are many oscillators and sometimes they make quasi-periodic solutions via their interactions. Therefore, bifurcation analysis of quasi-periodic solution attracts the attention of many researchers. Bifurcation point of quasi-periodic solution is detected by the point at which dominant Lyapunov exponent (abbr. DLE) touches zero. However, because we cannot apply classical Newton’s method for closed flow of periodic-solution to quasi-periodic solution, the analysis of bifurcation mechanism of quasi-periodic solution was difficult. For this reason, the analysis of quasi-periodic solution has not been much performed.

Recently, we have developed a bifurcation analysis tool for higher-dimensional torus named Lyapunov bundle (abbr. LB) [1], [2]. The LB is a generalization of eigenvectors to quasi-periodic solutions and a set of tangent vectors, called Lyapunov vectors (abbr. LV), on every points in a torus. Local bifurcation type is clarified by the topology of LB around a local bifurcation point.

2. Lyapunov Bundles on Section Tori

On an \( n \)-dimensional torus, there are \( n \) LBs and related \( n \) DLEs on the torus. Important factors are DLE and related dominant LB (abbr. DLB). In bifurcation diagram, we can find local bifurcation point by DLE. We can classify the type of local bifurcation by the topology of DLB. Table 1 shows relationship between local bifurcation types and DLB types on the original torus, ST1, and ST0. Fig. 1 represents schematics of DLB topology on section tori and local bifurcations. The shape of type-\( A^{n+} \)
DLB is $n$-dimensional orientation-preserving annulus. On 1-dimensional section, $A^{p+}$ is projected to 1-dimensional orientation-preserving one-sided annulus $A^+$. Furthermore, on 0-dimensional section (just a point), it is projected to 0-dimensional orientation-preserving one-sided annulus $A^{0+}$ which is one-sided vector. The shape of type-$A^{p+}$ DLB is $n$-dimensional orientation-reversing two-sided annulus. On 0-dimensional section, it is projected to 0-dimensional orientation-reversing two-sided annulus $A^{0+}$ which is two-sided vector. The shape of type-$M^n$ DLB is $n$-dimensional Möbius band. On 0-dimensional section, it is projected to 0-dimensional two-sided annulus shape $A^{b+}$. The topologies of $A^{b+}$ and $A^{0+}$ are the same. Nevertheless, we can distinguish them by taking twice-iterated map. Type-$F^n$ DLB is related to multiplicity-2 DLE and $n$-dimensional focus shape. If $n = 1$, it has test tube brush like shape. On 0-dimensional section, it is projected to 0-dimensional focus type DLB $F^0$ which has the shape of disc.

As a result of DLB calculation on 0-dimensional section, we can classify the bifurcation type on the original $n$-dimensional torus following 4 cases: Case 1: we obtained the $A^{p+}$-type DLB, local bifurcation of the original $n$-dimensional torus is tangent bifurcation (saddle-node, pitchfork, transcritical, etc.). Case 2: we obtained the $A^{0+}$-type DLB, local bifurcation of the original is period doubling bifurcation. Case 3: we obtained the $A^{b+}$-type DLB, local bifurcation of the original is double covering (DC) bifurcation. Case 4: we obtained the $F^0$-type DLB, local bifurcation of the original is Neimark-Sacker bifurcation.

3. Demonstration of DLB on ST0

For an example of local bifurcation analysis by DLB, we take 3rd-order Duffing equation with a periodic external force [3]. The normalized system equation is as follows:

$$\begin{cases} x_0 &= x_1 \\ x_1 &= -k_1 x_1 - 1/8(x_0^2 + 3x_1^2)x_0 + B \cos t \\ x_2 &= -1/8k_2(3x_0^2 + x_1^2)x_2 + B_0. \end{cases}$$

DC bifurcations of 2-torus occurs in this system. The bifurcation parameter is $B$ and the other parameters are fixed as follows: $k_1 = 0.03, k_2 = 0.05$, and $B_0 = 0.03$. Figure 2 represents FT2 attractor with one-turn before the DC bifurcation. Figure 3 represents FT2 attractor with two-turns after the DC bifurcation.
which has two-sided vector. From this result, we can eluci-
date that the local bifurcation of the original 2-torus attrac-
tor is DC bifurcation.

### 4. conclusion

We developed a bifurcation analysis tool “dominant Lyapunov bundle on 0-dimensional section torus” for analyz-
ing local bifurcation of high-dimensional torus which has complex shape. By using this DLB on ST0, one can eluci-
date the local bifurcation of the original heigh-dimensional torus easily. We demonstrated the DLB on ST0 for the dou-
ble covering bifurcation of complex 2-dimensional torus with winding in 3rd-order Duffing equation with a periodic external force. This DLB is applicable for any complex torus attractor, because this method uses just a point ST0 with small radius.

### Acknowledgements

The work of the first, second and fourth authors (K.K., M.K. and K.A.) were supported by JST through its Im-

**References**


Partial Chaotic Synchronization of Coupled Nonidentical Augmented Lorenz Oscillators

Kenichiro Cho† and Takaya Miyano‡

†Department of Mechanical Engineering, Ritsumeikan University
1–1 Noji-Higashi, Kusatsu, Siga, 525-8577 Japan
Email: kcho@fc.ritsumei.ac.jp, tmiyano@se.ritsumei.ac.jp

Abstract—Recently, we have developed the augmented Lorenz equations, which are expressed as a star network of \( N \) Lorenz subsystems sharing the scalar variable \( X \) as the central node and are specified by an \( N \)-dimensional real diagonal matrix \( M \), where \( M_j \) is always set to \( M_j = 1 \) and \( M_n \) with \( n \) running from 2 to \( N \) randomly takes \( n \) or \( n + 0.5 \). Here, we show the synchronizability of coupled nonidentical augmented Lorenz oscillators each specified by different \( M \) and discuss the dynamical properties in terms of the trapdoor one-way function often used in the existing cryptosystems for digital signature.

1. Introduction

Recently, we have proposed a chaotic cryptographic method using the augmented Lorenz equations [1], [2]. The augmented Lorenz model is a system of \( 2N + 1 \)-dimensional ordinary differential equations. In our chaotic cryptographic method, a message sender (Alice) and receiver (Bob) have identical augmented Lorenz oscillators specified by the same \( N \)-dimensional real diagonal matrix, denoted as \( M \). Here, \( M \) is used as a secret-key. Then, a ciphertext masked by the chaotic time series generated by Alice’s augmented Lorenz oscillator is decrypted using the same chaotic time series generated by Bob’s oscillator. The initial conditions and bifurcation parameters are shared by Alice and Bob as public information. When \( N = 101 \), for instance, the size of the secret key space amounts to \( 2^{51} \times 2^{100} \approx O(10^{30}) \), which is prohibitively large for an eavesdropper, called Eve, to break the secret key by a brute force attack. Our cryptographic method is different from Cuomo’s method [3], [4] in that our method does not resort to chaotic synchronization.

Perfect synchronization of coupled identical Lorenz oscillators was found by Pecora and Carroll [5], [6]. The augmented Lorenz oscillators can achieve perfect synchronization in much the same way as coupled Lorenz oscillators, since the augmented Lorenz equations inherit the dynamical nature of the Lorenz equations. As will be shown in this report, however, nonidentical augmented Lorenz oscillators exhibit partial synchrony when they are coupled in a particular way, i.e., the direct coupling by sharing the weighted sum of the variables of \( Y_n \), denoted as \( Y_{tr} \).

Public key cryptography such as RSA (Rivest-Shamir-Adleman) resorts to the trapdoor one-way function. RSA is a standard public cryptographic method and digital signature. One-way function is not reversible, whereas the trapdoor one-way function is reversible when one has the key.

As will be shown in this study, the synchronous behavior of coupled augmented Lorenz oscillators can provide the effect of the trapdoor one-way function. In this paper, we show the partial chaotic synchronization of the augmented Lorenz oscillators and discuss their dynamical properties in terms of the trapdoor one-way function.

2. Augmented Lorenz Model and Partial Chaotic Synchronization

2.1. Augmented Lorenz Model

The augmented Lorenz equations are defined as

\[
\begin{align*}
\frac{dX}{dt} &= \sigma[Y_{tr} - X], \\
\frac{dN}{dt} &= RX - MZX - Y, \\
\frac{dZ}{dt} &= MYX - Z, \\
Y_{tr} &= tr\left(\left[M^{-1}\right]Y\right),
\end{align*}
\]

where \( X \) is a scalar variable, \( Y \) and \( Z \) are \( N \times N \) diagonal matrices whose diagonal components are denoted as \( Y_n \) and \( Z_n \) (\( n = 1,\ldots,N \)), \( t \) is dimensionless time, \( tr() \) denotes the diagonal sum of a matrix, \( \sigma \) and \( R_0 \) are bifurcation parameters. The matrix \( R \) is defined using

\[
M = \text{diag}(M_1, M_2, \ldots, M_N),
\]

\[
W = \text{diag}(\sin \phi, \sin 2 \phi, \ldots, \sin N \phi),
\]

\[
\Phi = \text{diag}\left(\frac{\phi - \frac{1}{2} \sin 2 \phi}{M_N - 1}, \ldots, \frac{1}{M_N - 1} \sin (M_N - 1) \phi, \frac{1}{M_N + 1} \sin (M_N + 1) \phi\right).
\]
where $\phi$ is a bifurcation parameter. In this study, $M_1$ is always set to $M_1 = 1$ and $M_n$ with $n$ running from 2 to $N$ randomly takes $n$ or $n + 0.5$.

2.1. Partial Chaotic Synchronization

The augmented Lorenz oscillators can be coupled by sharing $Y_n$ between the oscillators, which is referred to as direct coupling in this paper. We consider a drive-response system with the direct coupling via $Y_n$. The drive system is subject to Eqs. (1), and the response system is subject to

$$\begin{align*}
\frac{dS}{dt} &= \sigma[Y_n - S], \\
\frac{dT}{dt} &= R'S - M'US - T, \\
\frac{dU}{dt} &= M'TS - U,
\end{align*} \tag{2}$$

where $S$ is a scalar variable, $T$ and $U$ are $N \times N$ diagonal matrices whose diagonal components are denoted as $T_n$ and $U_n$, and $M \neq M'$. The drive system is not identical to the response one in the sense that $M \neq M'$.

We define the synchronization errors between the drive and response oscillators as

$$\begin{align*}
e_1 &= S - X, \\
e_2 &= T - Y, \\
e_3 &= U - Z, \tag{3}
\end{align*}$$

where $e_1$ is a scalar variable, $e_2$ and $e_3$ are $N \times N$ diagonal matrices whose diagonal components are denoted as $e_{2n}$ and $e_{3n}$ with running from 1 to $N$, respectively.

In the case of the direct coupling via $Y_n$, Eqs. (1) - Eqs. (3), we obtain

$$\begin{align*}
e_1' &= -\sigma e_1, \\
e_2' &= [(R' - M'U) - (R - MZ)]X - e_2, \\
e_3' &= (MTS - U) - (MYX - Z). \tag{4}
\end{align*}$$

Since $e_1 = \exp(-\sigma t)$, $e_1 \to 0$ as $t \to \infty$. Such asymptotic behavior is also obtained in the case of sharing $X$ by the drive and response systems. Thus, Eqs. (4) is rewritten as

$$\begin{align*}
S &= X, \\
e_2 &= [(R' - M'U) - (R - MZ)]X - e_2, \\
e_3 &= (MTS - U) - (MYX - Z), \tag{5}
\end{align*}$$

that is, for $n = 1, \ldots, N$,

$$\begin{align*}
e_1 &= 0, \\
e_{2n}' &= [R'_n - R_n - (M'_nU_n - M_nZ_n)]X - e_{2n}, \\
e_{3n}' &= (M'_nT_n - M_nY_n)X - e_{3n}. \tag{6}
\end{align*}$$

When $M_n = M'_n$, we obtain

$$\begin{align*}
e_1 &= 0, \\
e_{2n}' &= -M_n e_{3n}X - e_{2n}, \\
e_{3n}' &= M_n e_{2n}X - e_{3n}. \tag{7}
\end{align*}$$

Let us introduce the Lyapunov function $E_n$ defined as

$$E_n = \frac{1}{2} [e_1^2 + e_{2n}^2 + e_{3n}^2]. \tag{8}$$

From Eqs. (6), the derivative of $E_n$ with respect to dimensionless time is obtained as

$$\begin{align*}
\dot{E}_n &= e_1 e_1' + e_{2n} e_{2n}' + e_{3n} e_{3n}' \\
&= [R'_n - R_n - (M'_nU_n - M_nZ_n)]Xe_{2n} - e_{2n}^2 \\
&\quad + (M'_nT_n - M_nY_n)Xe_{3n} - e_{3n}^2. \tag{9}
\end{align*}$$

Since $M_n = M'_n$, we obtain

$$\dot{E}_n = -e_{2n}^2 - e_{3n}^2. \tag{10}$$

The fixed point at $e_1 = 0, e_{2n} = 0, e_{3n} = 0$ is asymptotically stable. Hence, the chaotic synchronization is asymptotically achieved when $M_n = M'_n$. In the case of $M_n \neq M'_n$, from Eqs. (9), the chaotic synchronization is not achieved. Hence, the coupled nonidentical augmented Lorenz oscillators coupled via the direct coupling of $Y_n$ or $X$ are partially synchronized.

3. Numerical Results and Discussion

In our numerical experiments, the bifurcation parameters $\sigma$, $R_0$, $\phi$ were fixed at $\sigma = 25$, $R_0 = 3185$, $\phi = 0.36$, respectively. $M_n$ and $M'_n$ for $n = 2, \ldots, N$ were randomly set to $n$ or $n + 0.5$. We numerically integrated (1), (2) by the fourth-order Runge-Kutta method with a time width of $4 \times 10^{-4}$ and $N = 101$. The first 125,000 data points were discarded to eliminate the initial transient part of the numerical solutions. Figures 1(a), (b) show the numerical solutions of $X$ and $S$ as functions of time, respectively. Similar results for $Y_1, Y_{10}, Y_{100}, T_{10}$ and $T_{100}$ are shown in Figs. 2(a), (b), Figs. 3(a), (b), and Figs. 4(a), (b), respectively.

From Figs. 1(a), (b) and Figs. 2(a), (b), the variables $X$ and $T$ are in synchronization, despite $M \neq M'$. Figures 3(a) and (b) also indicate the synchronization between $Y_{10}$ and $T_{100}$, for which $M_{10} = M'_{10}$. However, Figs. 4(a) and (b) indicate that $Y_{100}$ and $T_{100}$, for which $M_{100} = M'_{100}$, are out of synchrony. Hence, it is concluded that the drive and response system achieve partial chaotic synchronization, despite $M \neq M'$.

Partial chaotic synchronization may be viewed as the trapdoor one-way function. In fact, the time series $X$ can
be reproduced using the time series $Y_{tr}$ through the partial chaotic synchronization with Eqs. (2). In contrast, the time series $Y_{tr}$ cannot be generated using the time series $X$ by the partial chaotic synchronization with Eqs. (2) unless $M = M'$. When Alice and Bob set $M$ to $M_a$ and $M_b$, respectively, Alice numerically integrates Eqs. (1) and send her time series $Y_{tr}$ to Bob. Then, Bob calculates Eqs. (2) and uses Alice’s time series $X$ to establish partial chaotic synchronization, whereas he cannot convert his time series $X$ to Alice’s time series $Y_{tr}$. These dynamical properties suggest that the partial synchronization provides the effect of the trapdoor one-way function.

Figure 1: Time series (a) $X(t)$ and (b) $S(t)$.

Figure 2: Time series (a) $Y(t)$ and (b) $T(t)$.

Figure 3: Time series (a) $Y_{10}(t)$ and (b) $T_{10}(t)$. 
4. Conclusions

We have shown that the direct coupling via $Y_u$ can generate partial chaotic synchronization between two nonidentical augmented Lorenz oscillators. The partial synchrony provides the effect of the trapdoor one-way function, which is often used in the public-key cryptography. In a future paper, we will propose a method for digital signature based on the partial synchronization of coupled augmented Lorenz oscillators.

Acknowledgment

This study is supported by JSPS KAKENHI Grant Number JP15K00353.

References


Investigation of Attracting Force to Synchronization States on Coupled Oscillator System by Using Electric Power

Ryoji Fukumasa†, Masayuki Yamauchi‡ and Yoshifumi Nishio§

†Hiroshima Institute of Technology
2-1, Miyake, Saeki-ku, Hiroshima, 731–5193, Japan
Phone/FAX:+81-82-921-4248
E-mail: ma16009@cc.it-hiroshima.ac.jp

‡Tokushima University
2-1, Minamijyosanjima-cho, Tokushima, 770–8506, Japan
E-mail: nishio@ee.tokushima-u.ac.jp

Abstract

In this study, the attracting forces to a steady states are theoretically analyzed by using reactive powers in coupled oscillator system, and are obtained by using a simulation. The attracting forces are investigated in changing a coupling parameter by using our theoretical method.

1. Introduction

We can observe synchronization phenomena in our body. There are a synchronization of among pacemaker cells in our heart, and the pulse synchronization phenomenon can be observed in other nervous system, too. Therefore, we can say that the synchronization phenomenon is one of most important phenomena in this world. The synchronization phenomena can be observed on electric circuits. Many kinds of synchronization phenomena and characteristics of the phenomena were reported by many researchers[2]. The synchronization phenomena have been analyzed by many methods. The averaging method is often used when the synchronization phenomena are theoretically analyzed[2]. The method is hard to be used for transient state.

In this study, an attracting force, which two oscillators are attracted to a synchronization state, are investigated by using reactive powers of the system and simulations. Furthermore, the attracting force is investigated in changing a coupling parameter by using our theoretical method.

2. Circuit Model

Our circuit model is shown in Fig. 1. The van der Pol oscillators are coupled by inductor \( L_c \). An inductor and a capacitor of each van der Pol oscillator are shown as “\( L \)” and “\( C \)” respectively. Characteristic of nonlinear negative resistor of \( k \)-th oscillator is shown as \( f(v_k) \) in Eq. (1).

\[
f(v_k) = -g_1 v_k + g_3 v_k^3 \quad (k = 1 \text{ or } 2)
\]

(1)

Circuit equations of this circuit are normalized by using Eq. (2). The normalized equations are shown in Eq. (3).

\[
i = \sqrt{\frac{C g_1}{3 L g_3}} x_k, \quad v = \sqrt{\frac{g_1}{3 g_3}} y_k, \quad t = \sqrt{L C} \tau,
\]

(2)

\[
f_k = f(v_k) = -g_1 v_k + g_3 v_k^3
\]

(3)

\[
\frac{dx_k}{dt} = y_k,
\]

\[
\frac{dy_k}{dt} = -x_k + a(x_0 - 2x_k + x_0) + \varepsilon(y_k - \frac{1}{3}v_k^3).
\]

(3)

Where \( \alpha \) expresses a coupling parameter and \( \varepsilon \) shows nonlinearity of each oscillator.

2.1. Calculation of reactive power

An instantaneous electric power of each oscillators and an instantaneous electric power of each inductor between adjacent oscillators are calculated by which each oscillation wave shape is assumed as a sinusoidal wave. Each current \( x_k \) (\( k = 1 \) or 2) and each voltage \( y_k \) are assumed as Eq. (4). Amplitudes of two oscillators are assumed as same value. Angular frequencies of two oscillators are assumed as same value, too.

\[
x_k = X sin(\omega t + \theta_k)
\]

(4)

\[
y_k = \omega X sin(\omega t + \theta_k)
\]

(4)

<Instantaneous electric power of the inductor in each oscillator>

\[
P_{L_k} = \delta \frac{\varepsilon}{\varepsilon} x_k y_k
\]

(5)

<Instantaneous electric power of the capacitor in each oscillator>

\[
P_{C_k} = \delta \frac{\varepsilon}{\varepsilon} y_k \frac{dy_k}{dt}
\]

(6)

<Instantaneous electric power of the coupling inductor>

\[
P_{L_c(2,1)} = \frac{\alpha \delta}{\varepsilon} (y_1 - y_2)(x_1 - x_2)
\]

(7)
A normalized equation of a total instantaneous reactive power are assumed as Eq. (8).

\[
P_{\text{ rall }} = \sum_{k=1}^{2} \left( \frac{\alpha \delta}{\varepsilon} (y_1 - y_2)(x_1 - x_2) + \frac{\delta}{\varepsilon} y_k^2 \frac{dx_k}{dt} \right)
\]  

(8)

2.2. Derivation of angular frequency

In this system, we can consider that a power effect is best when a reactive power, which can be assumed as sum of instantaneous electric powers of \( L, C \) and \( L_C \), is zero. In other words, we can guess that this system is a steady state when the reactive power is zero. In this study, angular frequencies of the in-phase synchronizations and the anti-phase synchronizations are obtained when the reactive powers are zero.

\[
P_{\text{ rall }} = 0
\]  

(9)

We want to calculate Eq. (8), but we can’t calculate the equation without an oscillation angular frequency \( \omega \) and an amplitude \( X \). However, when \( \theta_1 \) and \( \theta_2 \) are zero or \( \theta_1 \) is 0 and \( \theta_2 \) is \( \pi \), \( \omega \) and \( \omega_1 \) can be obtained by using Eq. (9).

\begin{align*}
\text{Phase angles of the in-phase synchronization} & : \theta_k = 0 \quad (k = 1 \text{ or } 2) \\
\text{Phase angles of the anti-phase synchronization} & : \begin{cases} 
\theta_1 = 0 \\
\theta_2 = \pi 
\end{cases}
\end{align*}

(10) (11)

Firstly, an in-phase synchronization frequency \( \omega_m \) and an anti-phase synchronization frequency \( \omega_{\text{anti}} \) of this circuit are calculated by using the Eq. (9). A phase of each oscillator is set as \( \theta_k \).

\begin{align*}
\text{Phase angles of the in-phase synchronization} & : \theta_k = 0 \quad (k = 1 \text{ or } 2) \\
\text{Phase angles of the anti-phase synchronization} & : \begin{cases} 
\theta_1 = 0 \\
\theta_2 = \pi 
\end{cases}
\end{align*}

(12) (13)

The angular frequency of the in-phase synchronization is achieved by which Eq. (12) is applied to Eq. (9).

\[

\omega_m = 1
\]  

(14)

The angular frequency of the anti-phase synchronization is achieved by which Eq. (13) is applied to Eq. (9).

\[

\omega_{\text{anti}} = \sqrt{1 + 2 \pi}
\]  

(15)

We assume that each angular frequency \( \omega \) of each phase difference \( \theta_2 - \theta_1 \)[degree] exists between \( \omega_m \) and \( \omega_{\text{anti}} \) and linearly vary between \( \omega_m \) and \( \omega_{\text{anti}} \). Therefore, \( \omega \) is calculated by the following equation.

\[

\omega = \frac{\omega_{\text{anti}} - \omega_m}{180} \times (\theta_2 - \theta_1) + \omega_m
\]  

(16)

Next, amplitude \( X \) is calculated. A total active power of this circuit is obtained as a sum of powers of a nonlinear negative resistors in each oscillator(see Eq. (17)).

\[
P_{\text{ all }} = \frac{2}{\pi} \sum_{k=1}^{2} \left( -\frac{\delta}{\varepsilon} \frac{d^2 x_k}{d\tau^2} + \frac{\delta}{\varepsilon^2} \frac{d^4 x_k}{d\tau^4} \right)
\]  

(17)

An amplitude \( X \) of each oscillator is calculated, when \( P_{\text{ all }} \) is integrated in a period and the result is assumed zero as follows.

\[

\int_0^\tau P_{\text{ all}} d\tau = 0
\]  

(18)

We calculate \( P_{\text{ rall}} \) of each phase difference by using the \( \omega \) and the \( X \).

3. Attracting Force

The attracting force is considered by using reactive powers and simulations.

3.1. Analyzing method by using reactive powers

We set \( \theta = \theta_2 - \theta_1 \). Waveforms of reactive powers are shown in Figs. (2)–(6) in changing the phase difference \( \theta \). We can understand that the amplitude is zero when the \( \theta \) is 0 degrees or 180 degrees. In other words, if a condition of the system is a steady state, the amplitude of the reactive power is zero, and if the condition is not the steady state, the amplitude is not zero. Therefore, we think that the system becomes unstable when a value, which the reactive power is squared and integrated in a period, become large. The value is expressed as \( P_{\tau T} \).

\[
P_{\tau T} = \int_0^\tau \text{Pall}^2 d\tau
\]  

(19)

The \( P_{\tau T} \) is calculated in changing phase difference \( \theta \) and shown in Figs. (7) and (8). The Fig. (7) shows results of when the \( \alpha \) is set as 0.05 and the \( \varepsilon \) is 0.1, and the Fig. (8) shows results of when the \( \alpha \) is set as 0.1 and the \( \varepsilon \) is 0.1. We can understand that \( P_{\tau T} \) becomes a maximum value when \( \theta = 90 \) degrees. In under 90 degrees, when the \( \theta \) becomes small, \( P_{\tau T} \) becomes small. In over 90 degrees, when the \( \theta \) becomes large, \( P_{\tau T} \) becomes small. We assume that \( \theta \) changes so that the \( P_{\tau T} \) becomes small. In other words, the \( \theta \) is attracted to zero under 90 degrees, and \( \theta \) is attracted to 180 degrees over 90 degrees. If change of the \( P_{\tau T} \) is large when the \( \theta \) is a little changed, we think that the attracting force is large. Therefore, we calculate gradient values of the graph of the \( P_{\tau T} \). The gradient values are multiplied by \(-1\), because we want to show the results as negative values in a domain of which the \( \theta \) decreases to zero, and as positive values in a domain of which the \( \theta \) increases to 180 degrees. The inversion gradient values \( g \) are shown in Figs. (9) and (10)(see Eq. (20)).
We can think that the attracting force to stable state is strongest when an absolute value of the \( g \) is a maximum value. Therefore, we can understand that when \( \theta = 45 \) degrees and 135 degrees attracting force become strongest.

### 3.2. Analyzing method by using simulations

An attracting force of each \( \theta \) is investigated by using a simulator. An initial phase difference between two oscillators is changed by which initial values of a voltage and
a current of each oscillator are changed. A calculation method of phase difference is shown as follows (see Fig. 11). In the Fig. 11, two sinusoidal waves show voltages of two oscillators. The $a_1$ shows the first positive peak of an oscillation waveform of an oscillator after $\tau = 0$, and the $a_2$ expresses the second positive peak. The $b_1$ shows the first positive peak of a oscillation waveform of another oscillator. Time of $a_1$, $a_2$ and $b_2$ are expressed $a_1$, $a_2$ and $b_2$.

The $\theta$ is calculated by using Eq. (21).

$$\theta = \frac{\tau_{b_2} - \tau_{a_2}}{\tau_{a_1} - \tau_{a_2}} \times 360[\text{degree}] \quad (21)$$

In this paper, we investigate how much the $\theta$ is changed during 1$\pi$ for each the initial phase difference. A changing value of $\theta$ is shown as $\Delta \theta$ and calculated by Eq. (22).

$$\Delta \theta = \left(\frac{\tau_{b_3} - \tau_{a_3}}{\tau_{a_2} - \tau_{a_1}} - \frac{\tau_{b_2} - \tau_{a_2}}{\tau_{a_2} - \tau_{a_1}}\right) \times \frac{1}{\tau_{a_2} - \tau_{a_1}} \times 360[\text{degree}] \quad (22)$$

The simulation results are shown in Figs. 12 and 13. The horizontal axes show the $\theta$, and the vertical axes show the $\Delta \theta$ in Figs. 12 and 13. Therefore, this graph means an attracting force $g$ which is strong when a absolute value of $\Delta \theta$ is large value.

We can observe that the shapes of graphs of simulation results are same shape with theoretical graphs, basically.

4. Comparison

Attracting forces are investigated in changing the coupling parameter $\alpha$ by our using theoretical method. The nonlinearity $\varepsilon$ is fixed as 0.1 and the $\alpha$ is set as 0.05, 0.10, 0.15 or 0.20. The theoretical results are shown in Fig. 14. We can understand that the attracting force $g$ becomes large as the coupling parameter $\alpha$ becomes large.

5. Conclusions

In this study, the attracting forces to a steady states were theoretically analyzed by using reactive powers and were obtained by a simulation in a coupled oscillator system. The theoretical results and the simulation results were observed same results, basically. Furthermore, we investigated that the attracting force becomes strong as coupling parameter is increased by using our theoretical method.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Number 25730152.

References


Clustered Multidimensional Scaling with Rulkov Neurons

Thomas Ott†, Martin Schüle‡, Jenny Held‡, Carlo Albert‡ and Ruedi Stoop*

†ZHAW Zurich University of Applied Sciences
8820 Wädenswil, Switzerland
‡EAWAG Swiss Federal Institute of Aquatic Science and Technology
8600 Dübendorf, Switzerland
* Institute of Neuroinformatics, ETH/UZH Zurich
Winterthurerstrasse 190, 8047 Zurich, Switzerland

Email: thomas.ott@zhaw.ch, scli@zhaw.ch, Jenny.Held@eawag.ch, Carlo.Albert@eawag.ch, ruedi@ini.phys.ethz.ch

Abstract—When dealing with high-dimensional measurements that often show non-linear characteristics at multiple scales, a need for unbiased and robust classification and interpretation techniques has emerged. Here, we present a method for mapping high-dimensional data onto low-dimensional spaces, allowing for a fast visual interpretation of the data. Classical approaches of dimensionality reduction attempt to preserve the geometry of the data. They often fail to correctly grasp cluster structures, for instance in high-dimensional situations, where distances between data points tend to become more similar. In order to cope with this clustering problem, we propose to combine classical multi-dimensional scaling with data clustering based on self-organization processes in neural networks, where the goal is to amplify rather than preserve local cluster structures. We find that applying dimensionality reduction techniques to the output of neural network based clustering not only allows for a convenient visual inspection, but also leads to further insights into the intra- and inter-cluster connectivity. We report on an implementation of the method with Rulkov-Hebbian-learning clustering and illustrate its suitability in comparison to traditional methods by means of an artificial dataset and a real world example.

1. Introduction

Visual inspection of scatterplots is a fast and common way to interpret data. Yet, high-dimensional data can be difficult to interpret. Hence, dimensionality reduction is often performed, aiming to achieve a compact representation of the data in two or three dimensions. A plethora of different techniques for dimensionality reduction has been proposed (see e.g. [1] for an overview). The most common methods are linear techniques such as principal component analysis PCA and multidimensional scaling (MDS). These methods often have difficulties with the representation of real world data, as in many cases high-dimensional data is generated by non-linear processes, resulting in highly non-trivial structures in the space of measurements [2]. Hence, during the last two decades, effort has been put in the development of nonlinear dimensionality reduction techniques in order to map nonlinear manifolds, e.g, kernel PCA, Isomap, Locally Linear Embedding, Diffusion Maps or t-SNE [3], to name but a few.

Typically, these approaches attempt to preserve the geometry of the data, at least at a local scale. That is, the distances between data points in a local neighborhood shall be preserved in the low-dimensional representation. Less attention has been paid to mapping cluster-like structures of a possibly complex (e.g. nonconvex) shape that can be intrinsically high-dimensional. Here, classical approaches often struggle because of unclear cluster boundaries that might even be obscured by data noise or due to the curse of dimensionality, notably the fact that distances between data points tend to become more equal with growing dimensionality. Real examples of such data situations are for instance encountered when dealing with flow cytometry data. In order to highlight the prevalent intrinsic cluster structures in the low-dimensional representation, we suggest to no longer stick to the goal of preserving the local geometry. Our idea is to combine self-organizing clustering processes with multi-dimensional scaling in order to enhance local cluster structures. To this end, we employ the Rulkov-Hebbian-learning clustering algorithm (RHLC) that has recently been introduced in [2] and can be considered an efficient cortex-inspired clustering method. The method presented here however also works with similar algorithms such as HLC with integrate-and-fire neurons [4]. In the following we briefly discuss the two ingredients of our approach, MDS and RHLC, and then present their application to two example data sets.

2. Multi-Dimensional Scaling and Extensions

We are given high-dimensional data vectors \( x_i \in \mathbb{R}^d \) \((i \in \{1, \ldots, n\})\) that form the rows of a matrix \( X \). The goal of classical multidimensional scaling MDS is to find low-dimensional (often 2-dimensional) reconstruction vectors \( y_i \) (or \( Y \) as a matrix) that minimize the following cost function [1]

\[
\Phi(y) = \sum_{i,j} (d_{ij} - \|y_i - y_j\|)^2 \tag{1}
\]
where \( d_{ij} = d_{ji} \) is the Euclidian distance between the original data points and \( ||y_i - y_j|| \) denotes the Euclidian distance in the reconstruction space. The minimum of (1) can be calculated based on the eigendecomposition of the Gram matrix \( K = XX^T \) which can be obtained by double-centering the distance matrix \( D = (d_{ij}) \) [1]. The \( m = 2 \) coordinates of \( y_i \) are then given by \( Y = E_m \Lambda_m^{1/2} \), where \( \Lambda_m \) is the diagonal matrix with the \( m \) largest eigenvalues and \( E_m \) is the matrix of the corresponding eigenvectors of \( K \). For Euclidian distances, principal component analysis (PCA) is identical to MDS due to the fact that the eigenvectors of the Gram Matrix and the eigenvectors of the covariance matrix are directly related. MDS can also be applied for any other distance matrix, a fact that is exploited by the Isomap approach with the goal to better account for describing the neighborhood of datapoints on curved manifolds. In Isomap, the distances are calculated by first constructing a graph, in which every point is connected to its \( k \) nearest neighbors. The distance between two points is then set to be the length of the shortest path in the graph [5].

In general metric MDS, the goal of (1) is relaxed by replacing \( d_{ij} \) with \( \delta_{ij} = f(d_{ij}) \), where \( f() \) leads to a new symmetric nonnegative (i.e. positive semi-definite) dissimilarity matrix. Isomap and similar methods can be interpreted as a kernel PCA method with the advantage of having a method that also works for datapoints that are not in the training sample.

3. From Rulkov Clustering to RHLC-MDS/Isomap

Rulkov-Hebbian-learning clustering (RHLC) is based on self-organization processes in a network of Rulkov neurons, letting clusters arise from the interplay between neural activity and changes in the network connectivity. Hebbian learning-based clustering (HLC) has been introduced in previous work as a remedy for the intrinsic shape biases introduced by standard clustering algorithms [2],[6]. Recently, RHLC has been developed, functioning on the same Hebbian-learning principle but making use of the more efficient map-based Rulkov neuron dynamics [7]. Generally, in HLC, every data item is interpreted as a dynamical unit with node dynamics, which are allowed to interact via a \( k \)-nearest neighbors graph. The pair-wise interaction strength between nodes is weighted so as to represent the local distances between the data items. In an iterative process, using Hebbian learning, the network structure is updated such that the weights between dynamically similar nodes are strengthened, while a counter-acting mechanism aiming to preserve the level of activity in the network causes the weights between less similar nodes to decrease. The final graph structure can thus be represented by the weights \( w_{ij}^\alpha \in [0,1] \), where strongly connected nodes will have a large coupling strength. The connectivity of nodes across a cluster can in this way easily be represented without the need for direct interaction, and thus without shape bias. For clustering, all weights below a threshold are deleted and the remaining (sub)graphs define the final clusters.

Here, for the purpose of data visualization, we are not using a hard cluster assignment. Rather, we are interested in the final weight matrix produced by RHLC as it reflects an amplified similarity between the data items. On the whole, RHLC performs a mapping

\[
d_{ij} \rightarrow \delta_{ij} = f(d_{ij}) = 1 - w_{ij}^\alpha,
\]

The matrix defined by \( \delta_{ij} \) can be interpreted as a ‘clustered distance matrix’ and can serve as input for MDS, giving rise to our RHLC-MDS method. As an alternative method, we use \( \delta_{ij} \) as input for Isomap. This RHLC-Isomap method is motivated by the observation that the basic connection matrix of RHLC reflects a \( k \) nearest neighbor graph as it is also used for Isomap.

4. Experiments and Results

Datasets: We test our methods on the basis of two datasets.

1. The first dataset is an artificial dataset in 3D, where two clusters cannot be discriminated by comparing inter- and intra-cluster distances. This is due to the fact that the inter-cluster distance in V3-direction is smaller than the clusters’ extension in V1- and V2-direction (see Fig.1 a)) and can be considered a low-dimensional simulation of the curse of dimensionality. The situation is reminiscent of many real-life data sets, where the data may stretch multiple scales in different dimensions. Additionally, the clusters are embedded into a noisy background of data points. The clusters are colored in Fig.1 b).

2. The second dataset contains 2443 flow cytometry measurements from 3 different phytoplankton species, reduced to 8 descriptors. The scatterplots for a pairwise selection of variables are shown in Fig.1 c), where the species clusters are color-coded.

Compared Methods and Evaluation: We compare 4 methods: standard MDS and Isomap and our clustered versions, RHLC-MDS and RHLC-Isomap. In order to assess the convenience of reading out the (by virtue of construction) expected clusters or classes from the 2D scatterplot we use the ratio of the mean inter-class distances and the mean of the distances between the expected classes A as an indicator, i.e.

\[
r = \frac{<d_{within}>}{<d_{between}>},
\]

where the distances are meant to be the Euclidian distances in the two-dimensional projections. For compact and clearly separated clusters, a small \( r \) is expected.

RHLC as well as Isomap involve parameters to be tuned. For the following, the tuning was made manually based on a visual inspection of the results. Hence, they reflect optimal solutions at a pragmatic level.
Figure 1: Datasets for evaluation. a)/b) artificial dataset, c) phytoplankton dataset. The expected clusters are color-coded.

Figure 2: Visualization results for 1) artificial dataset with a) MDS, b) Isomap, c) RHLC-MDS, d) RHLC-Isomap, 2) phytoplankton dataset with e) MDS, f) Isomap, g) RHLC-MDS, h) RHLC-Isomap

**Results:** The visualization results for both datasets and all the compared methods are subsumed in Fig. 2. We observe by first focussing at the results for the artificial dataset (Fig. 2 a)-d)) that normal MDS fails to display the clusters separately. While normal Isomap seems to do a satisfying mapping job regarding the internal structure of the data, the RHLC versions of MDS and Isomap both better display the cluster structure of the data. However, RHLC-MDS splits the black cluster into two subunits and, generally, tends to arrange the points in linear chains with an overall center of mass in the center of the coordinate system. For RHLC-MDS, the background points (noise) are concentrated in the center (Fig. 2 c)). In contrast, RHLC-Isomap separates the noise as an independent cluster (Fig. 2 d)).

Similar observations regarding the visual output characteristics of the methods can be made for the phytoplankton dataset (Fig. 2 e)-h)) with the initial difference that the clusters are more clearly separable and a noisy background is absent.

The evaluation using the indicator \( r \) confirms that RHLC-Isomap is clearly superior to the other methods (smallest \( r \)) regarding the capability of highlighting the overall cluster structures (Table 1). For RHLC-MDS the \( r \) measure indicates a performance that is in the range of normal Isomap,
but better than normal MDS. To some extent, the results can be explained by the chain-like shape of the clusters that results from RHLC-MDS. Take the example of the phytoplankton dataset: while the form makes it convenient for humans to grasp the group structure of the data, the inter-class distances become rather large and hence $r$ is increased. In the case of the artificial dataset, the aspect that RHLC-MDS splits one cluster into two units also leads to an increased $r$. At the same time, it hints at the existence of an internal cluster structure. In fact, a closer inspection of this cluster reveals two different areas, where the neighbor density of points reaches a maximum. These slight inhomogeneities are amplified by RHLC and are made visible as two branches in the RHLC-MDS plot. The observation is illustrated in Fig. 3. The neighbor density was defined as the number of neighbors within a ball of radius $R = 0.8$ divided by the volume of the ball. Fig. 3 b) shows the neighbor density as a function of the points of the split cluster in the $V_1 - V_2$-projection.

<table>
<thead>
<tr>
<th></th>
<th>artificial dataset</th>
<th>phytoplankton dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>MDS</td>
<td>4.00</td>
<td>1.91</td>
</tr>
<tr>
<td>Isomap</td>
<td>1.07</td>
<td>0.70</td>
</tr>
<tr>
<td>RHLC-MDS</td>
<td>1.32</td>
<td>0.49</td>
</tr>
<tr>
<td>RHLC-Isomap</td>
<td>0.246</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Table 1: $r$ values for artificial dataset and phytoplankton dataset.

Figure 3: The two branches in the RHLC-MDS plot a) reflect regions of high neighbor density $n$ in the corresponding cluster.

5. Conclusion and Outlook

Neural network based clustering algorithms such as RHLC allow for an unbiased detection of local cluster structures on the basis of self-organization. In this process, the neighborhood structure of the data is encoded as a weighted network that evolves in such a way that inhomogeneities are amplified. Hence, the emergence of clear cluster structures is possible even in cases, where the detection of clusters is very challenging, e.g. when facing high-dimensional measurements with non-linear cluster characteristics. Here, we demonstrated that dimensionality reduction techniques such as MDS and Isomap allow for a low-dimensional representation of the evolved clustering network, shedding light on both the intra- and the inter-cluster structure. While RHLC-Isomap separates clusters more clearly, RHCL-MDS elucidates the internal structure of clusters. This also allows for a more robust determination of the most natural number of clusters by means of a quick visual inspection.

The method employs clustering as a preprocessing step for a dimensionality reduction (DR) step, which switches the role of the steps in comparison to the standard data analysis procedure. Thus, for future research, the results suggest a clustering method that iteratively applies DR and clustering techniques. Alternatively, DR and self-organized clustering can be hybridized in one method, following the idea developed in [8].

References

Nonparametric clustering approach towards big data

Tom Lorimer†, Jenny Held‡, Carlo Albert‡ and Ruedi Stoop†∗
†Institute of Neuroinformatics and Institute of Computational Science
University of Zurich and ETH Zurich, Winterthurerstrasse 190, 8057 Switzerland
‡Eawag, Überlandstrasse 133, 8600 Switzerland
∗Email: ruedi@ini.ethz.ch

Abstract—Clustering in bioinformatics is a fundamental process involving computational issues that are far from being resolved. In our work, we propose a new approach to this problem and show preliminary comparisons to current leading methods in the field.

1. Introduction

The big data problem has infiltrated many areas of science, notably bioinformatics [1]. We focus on a specific bioinformatics problem here: the identification, discovery and interrelation of cell types. This problem has developed over recent decades into analysing automated simultaneous measurements of the abundance of tens of marker proteins on (or in) tens to hundreds of thousands of cells, most recently using mass cytometry [2]. This shift from individual to population level investigation gives rise to a new kind of difficulty in interpretation: how can structure be identified in a high dimensional space without introducing bias? It has long been known [3] that nonlinear systems give rise to convex-concave ‘clusters’ of similar systems (e.g. systems showing the same periodicity lie on shrimp-shaped domains in parameter space), and this has recently been suggested to manifest also in the space of observable features more generally [4]. This implies that techniques used to identify high dimensional structure in mass cytometry data need to be able to deal with convex-concave clusters. The necessity of dealing with convex-concave clusters in mass cytometry data has also been identified recently, and a new clustering algorithm specifically proposed to deal with this problem [5]. This work will discuss our preliminary investigation of this algorithm, and compare it to our own clustering approach.

2. Toward unbiased clustering

Standard clustering approaches have a cluster shape bias that precludes accurate clustering of convex-concave sets. This bias arises from a (sometimes implicit) non-local distance criterion, where the distance from a point to a set is used to define clusters [4]. In order to cluster data without introducing bias, we need to use purely local pairwise distances between points, but still somehow ‘integrate’ this information to the level of a set. As a solution to this problem, Hebbian Learning Clustering (HLC) has been proposed in a previous work [4, 6]. HLC ascribes a local ‘node’ dynamics to each data point, and allows the dynamics of the nodes to interact via a k nearest neighbours graph. The strength of interaction across each link in the graph is weighted according to the distance between the points it connects. By exploiting a very general trade-off between the similarity of the node dynamics (homeophily), and the level of activity in the network (homeostasis), the graph’s weights can self-organise in an iterative manner such that the final connectivity strength of the graph determines the clusters, without requiring direct interaction across the set, and thus without introducing cluster shape bias [4, 6, 7]. HLC has recently been updated to use a more flexible and efficient map-based node dynamics defined by the Rulkov neuron model [8], and to fully exploit the sparse connectivity of the k nearest neighbour interaction matrix, rendering this approach feasible for big data problems [1]. This latest version of our algorithm, Rulkov HLC (RHLC) is used in this paper, and is described in Ref. [1].

3. Current leading approaches in mass cytometry data analysis

3.1. Visualisation: t-SNE

Student t-distributed Stochastic Neighborhood Embedding (t-SNE) [9] is a dimensionality reduction algorithm created for the visualisation of high dimensional datasets. Recently, it has been adopted in flow- and mass-cytometry data analysis under the name viSNE as an interpretation aid [10]. t-SNE achieves this dimensionality reduction by trying to match the pairwise distances between the points in the high and low dimensional spaces, where each distance is represented by a weight. Without going into detail, we note three features of this process that may cause problems for the representation of high dimensional complicated convex-concave datasets: i) the weights in the high dimensional space are normalised locally about each point, thereby removing local point density information; ii) the weight between each pair of points is made symmetrical by taking the average, thereby introducing artificial inhomogeneity into the local distance information; iii) the weights in the high dimensional space are defined according to a Gaussian distribution, whereas those in the low dimensional space are defined according to a Student t-
distribution (with power law tails) resulting in a diminished sensitivity to the position of widely spaced points in the low dimensional space.

3.2. Clustering: PhenoGraph

Many clustering algorithms are currently in use on mass cytometry data (see e.g. [2, 12] for overviews). PhenoGraph [5] stands out in particular both for its claimed effectiveness and for the apparent similarity of its methodology to our own clustering algorithm that has been proposed to overcome the difficulties of standard approaches [4, 6]. PhenoGraph begins by constructing a weighted $k$ nearest neighbours graph between the input data points. There are however two clear points of difference from HLC: i) the weights of the graph are not determined using the Euclidean distance directly, but instead using the Jaccard distance calculated on the neighbourhood overlap of the points; ii) the subdivision of the weighted graph into clusters is achieved using a well-known community detection algorithm [13]. The PhenoGraph approach has been shown to produce results that are consistent with major features identified by manual analysis of mass cytometry data [5]. The manual analysis of mass cytometry data however, has a number of limitations. Analysis proceeds by ‘manual gating’: defining clusters by sequentially selecting the points within regions (‘gates’) in a succession of two dimensional projections of the data. Each individual gate may not necessarily be convex in the projection in which it is defined, but in all other dimensions it is convex. This places severe restrictions on the kinds of high dimensional data structures that can be identified by manual gating, and so the suitability of manual gating for high dimensional mass cytometry data has been repeatedly questioned (e.g. [5, 10]). Moreover, it is possible that either the Jaccard distance (which has a normalising effect on data density), or the objective function used in community detection algorithms (which compares weights within the entire community/cluster) might compromise the local information of the $k$ nearest neighbours graph and lead to a cluster size or shape bias.

4. Benchmarking

We test PhenoGraph by benchmarking against custom synthetic datasets. Clustering can be seen as an ‘unsupervised’ task: one does not know the ‘correct’ answer. Therefore, before clustering algorithms can be used on real data, they need to be shown to produce stable and accurate results over a wide range of parameters on suitable test data. We base our benchmarking on synthetic two dimensional datasets, as this simplifies both the specification of arbitrarily complicated data structures, and the detailed interpretation of clustering results far beyond what is possible using standard clustering quality measures. As interesting problem settings are high dimensional, we generate high dimensional test datasets by embedding two dimensional datasets in higher dimensions. This approach permits the detailed analysis of high dimensional clustering results in the original two dimensional space. Using synthetic data for benchmarking (as opposed to, e.g. manually gated mass cytometry data) moreover guarantees the accuracy of the test labels.

To provide an overview of the clustering results across a range of parameters, we use the standard $F_1$ score or $F$-measure, i.e. the harmonic mean between ‘precision’ and ‘recall’ of a given cluster $i$ with respect to a retrieved cluster $j$

$$F_{ij} = \frac{2f_p f_r}{f_p + f_r},$$ (1)

where the precision, $f_p$ is the fraction of points in the retrieved cluster $j$ that are correctly assigned to given cluster $i$, and the recall, $f_r$ is the fraction of the points in the given cluster $i$ that are assigned to the retrieved cluster $j$. For each given cluster $i$, $F_{ij}$ will be different for different $j$. We define $F_i = \max_j F_{ij}$, and as an overall characterisation of the clustering, take either a mean, giving the unweighted F-measure $F = \frac{1}{n} \sum_i F_i$, or a weighted mean, giving the weighted F-measure $F_w = \sum_i \frac{n_i}{N} F_i$, where $n$ is the number of given clusters $i$, and $N$ is the total number of data points. These are standard statistical measures used for the assessment of clustering algorithms in general, including cytometry clustering algorithms [11, 12].

4.1. Two dimensions

We generated a suite of datasets of varying difficulty, each containing convex-concave shapes, with varying degrees of background noise. Selected here as an illustrative (rather than representative) example is a dataset of two concentric rings, with equal uniform density, separated by a thin band of lower density uniform noise (when calculating the F-measure, the assignment of points in the band of low density noise was ignored). We see in Fig. 1 that PhenoGraph does not successfully cluster this dataset for any tested value of $k$. Despite claims to the contrary, there is a clear cluster shape/size bias that precludes the inclusion of the entire outer ring in one cluster before the inner ring is also included. RHLC, by contrast, can successfully deal with this problem for a wide range of parameters (Fig. 2).

4.2. Higher dimensions

Our high dimensional test dataset with convex-concave structures for benchmarking, is composed from test datasets in 2 dimensions of differing sizes and densities highlighting a range of different difficulties that may be faced in clustering natural data. We transformed this two dimensional composed dataset into 8 dimensions according to

$$(x, y) \rightarrow (x + y, x - y, x^2, y^2, xy, x^2y, xy^2, x^3y^2),$$ (2)
such that the original 2 dimensional dataset now sits on a 2 dimensional sub-manifold of an 8 dimensional space. Although the first two dimensions of the transformation simply apply a rotation to the original dataset (so that there exists a projection that retains the original structure), this a priori knowledge is not available to the algorithms we test.

To illustrate the difficulty of reverse transforming such convex-concave data from a high dimensional space to two dimensions, even in the case where they are known to lie on a 2 dimensional sub-manifold, we performed a t-SNE transformation of our high dimensional test data set [9]. While it is not to be expected that t-SNE should return the original 2 dimensional configuration of points, we found that the t-SNE transformed data could not reasonably be interpreted in a way that would return the correct point labeling. While the major convex sets were preserved, the major convex-concave sets were partitioned in such a way that the pieces were no longer adjacent in the two dimensional space. More complicated convex-concave sets were partitioned into many pieces spread across the two dimensional plane, illustrating the difficulty of using t-SNE transformed data for interpretation.

Testing PhenoGraph on our high dimensional dataset, we found that it suffered similar problems to the two dimensional case, namely, an inherent cluster size/shape bias as a function of its parameter (Fig. 3). Although the weighted F-measure appears to monotonically increase across the tested range, we observe that this already coincides with an incorrect coarse grouping and splitting of clusters that can be expected to deteriorate further with further increasing $k$.

RHLC avoids this inherent cluster size/shape bias, and can successfully cluster the data over a wide range of parameters (Fig. 4). Even for RHLC however, this dataset is exceptionally difficult. RHLC has no local density normalisation, and we note that it struggles to cluster the lowest density cluster. This points the way toward a sequential clustering approach in future implementations.
5. Real data and outlook

The synthetic data examples presented so far provide important insight into the limitations of PhenoGraph and t-SNE, but how do these manifest when they are applied to real data? We are currently testing RHLC on one of the datasets used to benchmark PhenoGraph: a mass cytometry dataset of healthy human bone marrow cells described in Ref. [14]. We find that RHLC consistently merges some large clusters that were split both by manual gating and PhenoGraph. However based on our synthetic results, where PhenoGraph made artificial partitions of the clusters, we are currently investigating whether this is an RHLC clustering error, or whether these groups of cell types are actually joined in this dataset in a continuum of cell differentiation in high dimensions.

Acknowledgments

This work was supported by SNF grant numbers CR32I3 159660 to C.A. and 200021-153542/1 to R.S..

References

Architectural Approach on Approximate Computing for Media Processing

Yuko Hara-Azumi†‡, Hisashi Osawa‡, Tanvir Ahmed‡

†Tokyo Institute of Technology, Japan  ‡JST PRESTO
Email: {hisashi, tanvira, hara}@cad.ce.titech.ac.jp

Abstract—This talk presents our recent work on approximate computing-aware architecture. Leveraging the fact that some emerging applications are tolerant to computational errors to some extent, our work employs the new computational paradigm, Approximate Computing, which aggressively lessens computational complexity by adopting approximation, at the architectural level for the sake of performance improvement and energy reduction. In this talk, we show our approach of extending existing data management techniques for media processing applications.
Highly nonrandom synaptic connectivity
and spontaneous ongoing fluctuation in cortical networks

Jun-nosuke Teramae†,‡

†Graduate School of Information Science and Technology, Osaka University
1–5 Yamadaoka, Suita, Osaka 565-0871, Japan
‡Center for information and neural networks, Osaka University
1–4 Gakuendai, Yamadaoka, Suita, Osaka 565-0871, Japan
Email: teramae@ist.osaka-u.ac.jp

Abstract—Synaptic connectivity of local cortical circuit is highly nonrandom. Connection strength among neurons in the cortex is not distributed on normally or on the Gaussian distribution. Rather, it is well described with a long-tailed, typically the lognormal distribution. The skewed distribution indispensably contributes to robust generation of spontaneous ongoing fluctuation in cortical circuit, which is now thought as a key feature to realize flexible and high performance stochastic computation in the brain. In this study, introducing recent results of studies about synaptic connections in cortical circuit, we provide a mathematical background of fluctuation generation owing to the lognormal distribution.

1. Introduction

Synaptic connections among neurons in local cortical circuit are neither homogeneous nor random [1]. Strength of connections, or amplitude of excitatory postsynaptic potential (EPSP) is distributed with a long-tailed distribution, typically the lognormal distribution [1-3]. The lognormal distribution of connection strength means that while almost all connections are weak, a few connections are extremely strong. Actually, while typical amplitude of EPSPs are less than 1 mV, amplitude of a few connections reach to about 10 mV that is dozens times larger than the typical value of amplitude.

Sparse strong connections are not randomly distributed in the cortical circuit. Physiological experiments reported that strong connections form “cluster” with significantly higher probability from naively expected from average connection rate of the network [1]. For pairs of neurons, if one of neurons receives synaptic input from another with large amplitude, probability of existence of connection with opposite direction is significantly high. Also, correlation between connection strengths of forward and backward connections for pair of neuron is high if exists. Similarly, for triplet neuron, the cluster coefficient of them, i.e. probability that two of them are connected when these two are connected to the other one, is significantly high for strong connections. It is also reported that the cluster structure closely related receptive field of neurons in primary sensory cortex [4].

Interestingly, the nonrandom features of synaptic connectivity affect crucially on spike dynamics of population of neurons in the network. Especially, recent theoretical studies revealed that highly heterogeneous distribution of EPSP and cluster structure of sparse strong connections are key to robustly generate spontaneous and ongoing fluctuation of cortical networks [5-8], which is actually observed in vivo and in vitro experiment and known as a ground state of cortex [9, 10]. Robustly generated fluctuation in cortex is now thought as an indispensable factor of cortical computation. Accompanied with nonlinearity of neurons and population dynamics of neurons, the intrinsic fluctuation adjusts neural response and allows the network to have rich variety of states and responses.

In this study, we introduce recent developments of synaptic connectivity in local cortical circuit including cluster structure of them. Then develop a mathematical framework to characterize synaptic communication with the lognormal strength distribution. Owing to the long-tail nature, synaptic interactions on the network give very differently outcomes from normally connected network with for instance with the Gaussian distribution. We discuss how the specific feature of the lognormal distribution relates intrinsic fluctuation and computation in the brain.

Acknowledgments

This work is partially supported by Kakenhi 25430028 and JP16H01719.

References

A Study on Performance of Hopfield-Tank Neural Networks Running on Coherent Ising Machine

Hiroyuki Yasuda†, Mikio Hasegawa‡ and Kazuyuki Aihara†

†Institute of Industrial Science, The University of Tokyo
4–6–1 Komaba, Meguro-ku, Tokyo 345-0826, Japan
‡Department of Electrical Engineering, Tokyo University of Science
6–3–1 Niijuku, Katsushika, Tokyo 125-8585, Japan
Email: yasuda@sat.t.u-tokyo.ac.jp, hasegawa@ee.kagu.tus.ac.jp, aihara@sat.t.u-tokyo.ac.jp

Abstract—Coherent Ising Machine (CIM) which are composed with mutually interacted LASER network, and that have been studied as ultra-fast calculator for various optimization problem. It have been shown that CIM can obtain exact solution of MAX-CUT problem rapidly with high probability. We have proposed to use Hopfield-Tank Neural Network (HTNN) running on CIM to solve various optimization problems, such as traveling salesman problem. HTNN is one of combinatorial optimization method, and it have been shown that HTNN have ability to obtain exact solution on many kinds of combinatorial optimization problem and its real problem. In this paper, we have applied HTNN to CIM, and solved traveling salesman problem by HTNN running on CIM. We also investigate performance of HTNN running on CIM with using some kinds of noises.

1. Introduction

Coherent Ising Machine (CIM)[1] have been studied as one of combinatorial optimization method. CIM can obtain exact solution of several combinatorial optimization problem with high probability. For applying CIM to many kinds of combinatorial optimization problem, feasibility and implementation is studied. Especially, it have been shown that CIM can obtain exact solution of the max-cut problem[2] that is one of combinatorial optimization problem.

Combinatorial optimization method by CIM is based on the phenomenon that the energy of spin network minimize. Hopfield-Tank neural network (HTNN) is also mutually connected network, and have ability to solve combinatorial optimization problem such as traveling-salesman problem (TSP), quadratic assignment problem (QAP), and so on. Optimization method by HTNN is based on energy reducing in a monotone manner by network update, and which have disadvantage that the updating have tend to trap on local minimum. Therefore, it is necessary to improve the performance of HTNN by introducing stochastic fluctuation by Boltzman machine[4] or by using chaotic fluctuation with chaos neural network[5], but it is very hard to obtain exact solution even by these improving method.

In this paper, we propose fast optimization method using HTNN running on CIM which can obtain exact solution. First, we evaluate the feasibility of proposed method by applying the method to TSP. In this case, we introduce noise sequences to solve TSP by the HTNN model. Therefore, we investigate relationship between noise amplitude and performance of the model with changing the Gaussian sequences to other chaotic sequences.

2. Coherent Ising Machine

CIM is the system which can obtain ground state of Ising Hamiltonian. The state of spin glass is defined as following equations (1) and (2).

\[ \frac{dc_i}{dt} = (-1 + p - c_i^2 - s_i^2)c_i + \sum_{j=1}^{N} \xi_{ij}c_j, \] (1)

\[ \frac{ds_i}{dt} = (-1 - p - c_i^2 - s_i^2)s_i + \sum_{j=1}^{N} \xi_{ij}s_j. \] (2)

Where, \( c_i \) is normalized in-phase state, \( s_i \) is normalized quadrature phase state, \( p \) is pump rate, \( \xi_{ij} \) is mutual injection.

Each spin have binary states such as up and down. Total energy of mutually coupled spin network is expressed by following equation (3).

\[ H = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{N} A_i\sigma_i. \] (3)

Where, \( J_{ij} \) corresponds coupling strength, \( \sigma_i = \{ \pm 1 \} \) corresponds state of each spin.

Network becomes stable at ground state on appropriate pumping rate with updating the state of spins by equation (1) and (2). When CIM is used as optimization method, it is necessary to determine Hamiltonian to correspond with maximizing or minimizing objective function, and determine the connection weights between each spin based on the Hamiltonian. Hamiltonian becomes minimum state when CIM becomes ground state by updating, and that makes it possible to obtain solutions of minimum or maximum value of the objective function.
3. Hopfield-Tank Neural Network

Hopfield-Tank neural network is a mutually connected neural network. Updating of Hopfield-Tank neural network is expressed by following equation.

\[
\frac{dx_{ij}(t)}{dt} = -\frac{x_{ij}(t)}{\tau} + \sum_{k=1}^{N} \sum_{l=1}^{N} W_{ijkl} x_{kl}(t) - \theta_{ij}(t). \tag{4}
\]

Where, \(W_{ijkl}\) is coupling weight between neuron \(x_{ij}\) and \(x_{kl}\). \(W_{ijkl} = W_{ilkj} = 0\). \(\theta_{ij}\) means threshold of neuron \(x_{ij}\) and \(\tau\) means attenuation. \(X_{ij}(t)\) means output of neuron \(x_{ij}\), which have continuous range between 0 and 1. This output \(X_{ij}(t)\) is decided by following sigmoid function as output function of \(x_{ij}(t)\).

\[
X_{ij}(t) = \frac{1}{1 + \exp\left(-\frac{x_{ij}(t)}{\epsilon}\right)}. \tag{5}
\]

Where, \(\epsilon\) is parameter of sigmoid function. The energy of mutually coupled neural network by neuron of equation (4) is expressed by following equation (6).

\[
E(t) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{l=1}^{N} W_{ijkl} x_{ij}(t) x_{kl}(t) + \sum_{i=1}^{N} \sum_{j=1}^{N} \theta_{ij} x_{ij}(t). \tag{6}
\]

The energy of network is decreased according to neuron updating expressed by equation 4.

4. Optimizing Method of TSP by Hopfield-Tank Neural Network

Traveling salesman problem (TSP) is one of combinatorial optimization problems to find the shortest path to visit all destination. On TSP, \(i\) is index of city and \(j\) is order of visiting, and \(x_{ij} = 1\) means city \(i\) is visited at order \(j\). Total path length can be expressed \(\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} d_{ik} (x_{ik} + x_{kj})\). \(x_{ij}\) can be 1 for each \(i\) only once, because of that each city can be visited only once. Therefore, it is necessary to be \(\sum_{i=1}^{N} (\sum_{j=1}^{N} x_{ij} - 1)^2 = 0\).

Similarly, \(\sum_{j=1}^{N} (\sum_{i=1}^{N} x_{ij} - 1)^2\) should be 0. These are constraint function of TSP. By these path length and constraint functions, objective function of optimization on TSP can be expressed by following equation (7).

\[
E = A \sum_{i=1}^{N} \left(\sum_{j=1}^{N} x_{ij} - 1\right)^2 + B \sum_{i=1}^{N} \left(\sum_{j=1}^{N} x_{ij} - 1\right)^2 + C \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} d_{ik} (x_{ik} + x_{kj} - 1). \tag{7}
\]

Here, by comparison energy of HTNN and objective function of TSP, we can obtain injection weight \(W_{ijkl}\) and threshold \(\theta_{ij}\) as following:

\[
W_{ijkl} = -A\delta_{ij}(1-\delta_{ji}) - B\delta_{ij}(1-\delta_{ji}) - C\delta_{ik}(x_{ik} + x_{kj}), \tag{8}
\]

\[
\theta_{ij} = -(A + B). \tag{9}
\]

Energy function, objective function of TSP, converge to minimum solution by updating neuron state based on equation (4) with parameters \(W_{ijkl}\) and \(\theta_{ij}\).

5. Hopfield-Tank Neural Network Running on Coherent Ising Machine

In this paper, we investigate the performance of HTNN running on CIM. The output of HTNN expressed by the equation (5) have binary output 0 or 1. On the other hand, CIM spin that expressed by the equations (1) and (2) have binary output of –1 or 1. To homologize the output of HTNN and CIM, we have developed the equation of HTNN that have ±1 output. The output function of ±1 HTNN can be expressed by following equation (10) by assigning \(\hat{X}_i = 2X_i - 1\) to equation (4).

\[
\frac{dx_{ij}(t)}{dt} = -\frac{x_{ij}(t)}{\tau} + \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{W_{ijkl}}{2} \hat{X}_{kl}(t) - \left(\theta_{ij}(t) - \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{W_{ijkl}}{2}\right). \tag{10}
\]

Connection weight and threshold of HTNN can be decided by equation (10), and expressed by following equation (11).

\[
W_{ijkl} = \frac{W_{ijkl}}{2}, \quad \hat{\theta}_{ij} = \theta_{ij} - \sum_{k=1}^{N} \sum_{l=1}^{N} \frac{W_{ijkl}}{2}. \tag{11}
\]

Deeming \(c_{ij} > 0\) as firing of neuron, we obtain following equation (12).

\[
\frac{dc_{ij}}{dt} = (-1 + p - c^2_{ij} - \bar{s}_{ij}^2)c_{ij} + \hat{W}_{ijkl} c_{ij} + T_{scale}\hat{\theta}_{ij}. \tag{12}
\]

By this equation (12), it is made possible to run HTNN on CIM.

To investigate the performance of this new model, we try to obtain optimal solution of TSP 10 city problem by proposed method. Figure 1 shows rate of obtaining optimal solution with changing pump rate and noise amplitude. From figure 1, we can find that there are parameters that HTNN can obtain optimal solution. Therefore, we have found that HTNN running on CIM is feasible to solve combinatorial optimization problem such as TSP.
Figure 1: Rate of obtaining optimal solution with changing pump rate and noise amplitude using Hopfield-Tank neural network running on CIM.

Figure 2: Relationship between achieving rate of obtaining optimal solution and noise amplitude.

We have also confirmed whether performance of HTNN is improved by noise amplitude and noise sequence. Figure 2 shows relationship between achieving rate of obtaining optimal solution and noise amplitude. From figure 2, it can be found that there is appropriate noise amplitude to improve performance of HTNN. We also introduced logistic equation that have negative autocorrelation as noise sequence to HTNN. It have been found that solution searching performance of HTNN is improved slightly by logistic sequence. To improve the performance, it is also necessary to consider noise sequence that can be introduced to real machine of CIM.

6. Conclusion

In this paper, we have compared output function of CIM and HTNN and homologized the outputs of two models. By the proposed method, HTNN have been able to run on CIM. We have solved 10 city problem of TSP with proposed update function, and found that HTNN running on CIM have the ability to search the optimal solution.

In this paper, we have shown one example of methods to solve various combinatorial optimization problems. However, because real implementation of CIM is under considering, it is necessary to continue improving applicability of our proposed method. It may be necessary to modify equations to run HTNN on real machine of CIM. We also study other methods to extend applicable area of CIM on combinatorial optimization problems.

Acknowledgments

This work is supported by Japan Science and Technology Agency (JST) through its ImPACT program.

References


Distributed multipath routing with packet allocation based on the attractor renewal model

Yoshihiko Saitoh†Jun-nosuke Teramae †and Naoki Wakamiya‡

†Graduate School of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
Email: y-saitoh@ist.osaka-u.ac.jp, teramae@ist.osaka-u.ac.jp,wakamiya@ist.osaka-u.ac.jp

Abstract—Information communication network is rapidly growing recently. In order to reduce control overhead and realize efficient routing, distributed routing protocols attract much attention. One of these protocols is the attractor selection model that has been proposed inspired from a mechanism of gene expression in living cells. Distributed routing protocols including the attractor selection model, however, often fail to properly respond to traffic changes that occur on non-used paths because these protocols cannot aware status of network beyond their local scope. In order to overcome the problem, we propose a distributed multipath routing protocol based on ”attractor renewal model” that is an extension of the attractor selection model. In order to show validity of the proposed model, we perform network simulation where attractor selection model. In order to show validity of the proposed model, we perform network simulation where traffic condition changes in time. Results show that the proposed protocol surely rearranges allocation of packets depending on traffic condition beyond local scope of each node.

1. Introduction

Centralized routing protocols that requires information of whole network to operate now starts to face difficulties of robust and smooth operation due to increase in communication overhead accompanied with recent rapid growth of size and diversity of information communication networks [1]. In order to reduce communication overhead and realize robust and smooth routing that can be applicable even to much larger network, distributed and adaptive routing protocols in which operation on each node requires only local information of network such as communication delay on a few routes through the node attracts much attention recently.

These adaptive and distributed protocols include methods proposed inspired from adaptive nature of biological systems [3]- [7]. Biological systems often seem to have ability to promptly respond to environmental change even though the change is suddenly occurred. Moreover, these abilities often require only limited knowledge of their environment. Whereas these responses are not always optimal, quick, and often stochastic, responses of biological systems allow animals to survive in harsh environment with severe battles for existence.

The routing protocol based on the Adaptive Response Attractor Selection model (ARAS) is one of these biologically inspired protocols. ARAS was originally proposed to describe nonlinear dynamics of gene expression of a cell, E. coli. If the cell is in an environment with insufficient nutrient, gene expression network of the cell about metabotropic process changes itself to synthesize the depleted nutrient adaptively. Phase space of the dynamics has stable attractors corresponding to different nutrients and the adaptive dynamics is well described by autonomous selection of one of these attractors. Intriguingly, the dynamics is not deterministic. Rather, biological experiments and theoretical analysis of the phenomena reports that the process is stochastic and underlying fluctuation plays an essential role to adaptive attractor selection, which allows the network to suitably respond to even sudden change of environment [2].

Proposed routing protocol based on the ARAS assigns each attractor in phase space to a possible choice of next hop nodes on each node. Each node on the network calculates the stochastic dynamical equation independently. Depending on measured communication delay of a selected attractor, i.e. a selected next hop node, relative strength of fluctuation to deterministic force attracting the system to an attractor in the dynamical system changes adaptively. Strength of fluctuation increases, if the delay is large, and strong fluctuation forces the system to exit from the current attractor. On the contrary, if delay of the current selection decreases, relative strength of fluctuation decreases, which allows the system stays the current selection [5], [6].

Because the routing protocol can work without information of the whole network such as topology or connectivity of whole network or average delays on all pairs on nodes, the protocol largely reduces communication overhead comparing with centralized routing protocols as OSPF (Open Shortest Path First). This advantage allows us to set duty cycle of the protocol short and to realize packet routing that promptly reacts to rapid change of environment such as sudden change of traffic on the network.

The lack of global information, however, can be a significant drawback of the distributed routing protocol. Because attractor selection, or route selection, on each node is performed only based on communication delay along currently attractor, i.e. currently chosen next hop node,
the protocol cannot aware change of communication condition along currently non-used next hop node. Even though background traffic along a next-hop node decreases suddenly, for example, and communication delay along the next-hop node decreases largely, the ARAS cannot aware of appearance of the better route if ARAS does not select the route now. Due to random nature of the ARAS, it is still possible for ARAS to accidently find the better route. The response time, however, may not be sufficiently short.

In order to overcome the problem of the ARAS, here, we propose a distributed multipath routing protocol in which position of an attractor adaptively and continuously renewed based measured communication delay. On the contrary to the ARAS, the proposed protocol does not select a next-hop node. Rather, it adaptively changes “weights” of possible next-hop nodes based on position of an attractor and use these nodes with a ratio that proportional to the weight when it sends a packet.

In this paper, we first introduce the ARAS in the next section, and propose our routing protocol in section 3. We confirm validity of the proposed model using numerical simulation about queuing network in section 4. Conclusion is given in the last section.

2. Adaptive Response Attractor Selection model

2.1. Attractor selection model

Attractor selection model [2] is a nonlinear mathematical model describe the mechanism of E. coli cells adapt to surrounding nutrient environment by changing gene expression according to metabolic network, and synthesize lacking nutrient. In [2], genes which synthesize two different nutrients suppress the other gene expression and reach a stable state, called attractor, in which genes synthesize one of nutrients stably. The following equation describes concentrations of mRNA $\bar{m} = (m_1, m_2)$ corresponding to nutrient synthesis.

$$\frac{dm_1}{dt} = \frac{S(\alpha)}{1 + m_2^2} - D(\alpha)m_1 + \eta_1$$

$$\frac{dm_2}{dt} = \frac{S(\alpha)}{1 + m_1^2} - D(\alpha)m_2 + \eta_2.$$  \hspace{1cm} (1)

Here, $\alpha$ is activity of the cell, which represents goodness of current selection. $S(\alpha) = \frac{\alpha}{2m_0^2}$ and $D(\alpha) = \alpha$ represent functions of gene synthesis and degradation respectively, $\eta_1, \eta_2$ are fluctuation of gene expression, which is represented white Gaussian noise.

Figure 1 conceptually illustrates behavior of the attractor selection model to environmental change. When activity is high, potential of $\bar{m}$ is enough deep to E. coli stays stably in the current attractor even under fluctuation. When activity becomes low, potential of $\bar{m}$ becomes flat. Relative strength of fluctuation of gene expression becomes larger, and the system starts to exit from the current attractor to explore new attractor that provides higher activity. If the system successfully finds a good attractor with high activity, depth of potential of $\bar{m}$ increases again.

2.2. Routing with Attractor selection model

In [5], attractor selection model is extended to $M$-dimensional dynamical systems in order to apply $M$ candidates of next-hop.

$$\frac{dm_i}{dt} = \frac{s(\alpha)}{1 + m_{\max}^2 - m_i^2} - d(\alpha)m_i + \eta_i.$$ \hspace{1cm} (3)

$M$ is the number of candidates next hops of a node where the routing protocol is performed, $\alpha(0 \leq \alpha \leq 1)$ is the activity that represents goodness of current path defined as,

$$\alpha(h) = \frac{\min_{0 \leq k \leq W-1} [w(h - k)]}{w(h)},$$ \hspace{1cm} (4)

where $w(h)$ represents communication delay at time $h$, $W$ is a number of memory of past communication delay of used path. $m_{\max} = \max(m_1, \ldots, m_M)$, $s(\alpha) = \alpha(\beta e^\gamma + \phi')$, $d(\alpha) = \alpha$, $\phi' = \frac{1}{e\gamma}$, and $\eta_i$ is a noise term. Equation (3) has $M$ attractors, where one of $\bar{m}$, e.g. $m_j$, $j = (1, \ldots, M)$, takes a high value and the others take a low value. Each node identically calculates $\alpha$ and $m_i$ for all destination router.

Because the activity is defined as ratio between delay of current step ($h$) and the minimum delay over past $k$ steps, it decreases when communication delay get worse, which increases relative strength of the noise term and forces the protocol to find a better path. Note that if delay along currently selected path keeps constant value, the value of the activity continues to keep unity, which implies that the router continues to select the same path even though other better paths appear in currently non-selected paths.

2.3. Routing with Attractor renewal model

In order to solve the problem of limited scope and allow routers change their route to better paths even though these paths are not the currently used path, we propose a multipath routing protocol by extending attractor selection.
model. The routing protocol controls the ratio of packet allocation according to communication delay of path for effective use of network rather than select one route. We use $\bar{m}$ to represent the ratio of packet allocation to next-hop node as the ratio of the $i$th node $p_i$ is proportional with $m_i$. This means that position of attractor in phase space corresponds to ratio of packet allocation, and in order to control allocation adaptively, we need to change position of the attractor flexibly with reflecting goodness of these paths. As the control mechanism, we propose attractor renewal model, in which the position of attractor in phase space is updated with reflecting communication delay measured by using these routes.

We define function $f(d_i)$ that will be used to associate ratio of packet allocation to the $i$th next-hop nodes and communication delay measured for the node $d_i$ as,

$$f(d) = \max\{e^{-t_d^2}, \ C\}.$$  

$D$ defines the basis of communication delay and $C$ decides minimum value of $m_i$, which characterizes the lower bound of frequency in which $i$th next-hop node is selected. If the router cannot measure communication delay $d_i$ in some reason, $f(d_i)$ is set $C$.

Using values of $f(d_i)$ we define dynamics of $m_i$ as $m_i$ linearly decays to $f(d_i)$ with the time constant $\tau$. Finite nonzero value of $\tau$ contributes to prevent flapping of packet allocation.

$$\frac{dm_i}{dt} = -\frac{1}{\tau}(m_i - f(d_i)).$$  

(6)

After updating $\bar{m}$, $p_i$ is also updated as $p_i = m_i/\sum_{k=1}^{M} m_k$. We update these values with control period $T$.

3. Evaluation of the proposed protocol using network simulations

In this section, we shows validity of the proposal routing protocol with showing that the protocol can properly rearrange packet allocation when traffic along a next-hop node suddenly decrease even though the next-hop node was a suboptimal. Note that, original ARAS cannot notice the decrease of traffic along the next-hop node because the route from the next-hop node was suboptimal and does not used before the traffic decrease.

3.1. Simulation settings

We evaluate performance of the proposed protocol using a queuing network. We generate a random network using the Waxman model [8] of the number of nodes $N = 20$ and the number of edges $E = 30$. Capacity of all link is set to 100 Mbps and propagation delay along them is set to 3 msec. Packet size is fixed to 10000 bits, and $TTL$ is set to 15.

We set simulation time as 300 sec, and the control period $T$ as 1 sec. We also measure performance of the original ARAS with parameters that are set as given [5]. Parameters of the proposal model $\tau = 1$, $C = 0.001$, and $D = 5.0$.

3.2. Traffic variation

In this simulation, we set traffic of all pairs of node that connected each other as 334 kbps excepting one of the congested link that has 100 Mbps. At the time of 75 sec, we decrease traffic of the congested link to 334 kbps. Before 75 sec, both routing protocols, original ARAS and the proposed one, may avoid the congested link to send packets to their destination. After 75 sec, however, in order to decrease communication delay and efficiently use network resources, it must be preferable for routing protocols to allocate packets even to the previously congested link in addition to other links because the congestion has been resolved. Moreover if the congested link is included on the shortest path for some sessions, this session should use the link after 75 sec because the link must give the smallest communication delay.

3.3. Results of simulation

Figure 3 schematically shows behavior of the attractor renewal model. Depending on measured communication delay along $i$th next-hop node, value of $m_i$ changes to converge to $f(d_i)$, which implies that position of the attractor continuously moves.

Figure 2: In attractor renewal model, $\bar{m}$ can takes flexible value

Figure 2: In attractor renewal model, $\bar{m}$ can take flexible value.
cannot notice appearance of the optimal path that is different from the previously optimal path and cannot decrease the communication delay.

In order to show that above decrease of the communication delay is surely induced by packet allocation, we show time evolution of ratio of packet allocation to next-hop nodes in Figure 4. Before 75 sec, the node sends its packets to next-hop nodes 1, 3, and 5, and rarely sends its packets to the node 6 that gives shortest path because the path contains the congested link. After 75 sec, however, congestion is resolved and communication delay along the shortest path, 6, becomes the smallest value. As shown in the figure, accompanied with the improvement of communication condition along the route, ratio of packet allocation to the next-hop node 6 gradually increases, which results in better performance as shown in Figure 3.

4. Conclusion

In the paper, we have developed a novel adaptive and distributed routing protocol. Unlike the original ARAS, the proposed routing protocol simultaneously multiple paths to send their packet to their destination. Depending on measured communication delays of these paths, the protocol adaptively rearrange allocation ratio of packets. Because of parallel usage of multiple paths, the proposed algorithm can respond traffic change that occurs on previously sub-optimal paths, which cannot get noticed by the original ARAS. Using a network simulation with queuing network where traffic on them change in time, we show the proposed protocol surely rearrange traffic allocation and realize routing with smaller communication delay even while the change occurs along paths that are not optimal and rarely used previously.

Acknowledgments

This work was partially supported by Kakenhi 25430028 and JP16H01719.

References


**Effect of a Root-Raised-Cosine Filtered BPSK Signal on a Stochastic Resonance Receiver**

Yasuo Nakashima, Hiroya Tanaka, Takaya Yamazato, Yukihiro Tadokoro and Shintaro Arai

†Nagoya University, Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan
‡TOYOTA Central R&D Labs., Inc., 41-1 Yokomichi, Nagakute, Aichi, 480-1192, Japan
††Okayama University of Science, 1-1 Ridai-cho, Kita-ku, Okayama, 700-0005 Japan

Email:nakasima@katayama.nuee.nagoya-u.ac.jp

**Abstract**—Signal filtering is necessary for wireless communication however it fluctuates the signal amplitude and affects the performance of a stochastic resonance (SR) receiver. In this paper, we evaluate the bit error rate (BER) performance of filtered binary phase shift keying (BPSK) on an SR receiver. The result shows that filtering improves the BER performance of the SR receiver because the amplitude fluctuation contributes to improving the SR effect, indicating the applicability of the SR receiver to bandlimited BPSK signals.

1. Introduction

Stochastic Resonance (SR) is a nonlinear phenomenon that can enhance the response of a system by adding noise under certain conditions [1]. In contrast to many other systems that deal with noise negatively, SR positively utilizes noise. A particular advantage of SR is that it can detect weak signals buried in noise.

The characteristic of this interesting phenomenon has been discussed in the context of nonlinear physics [2,3], and some applications of SR to wireless communication have been proposed. SR is expected to be utilized for spectrum sensing in cognitive radio [4], signal detection [5,6], and improvement of receiver sensitivity [7].

We focus on applying SR to a receiver, for communication with weak signals. As application of SR to a receiver may have better sensitivity compared to a conventional receiver, and it can detect a significantly weak signal [8]. In previous research, an SR receiver in radio frequency (RF) has been proposed [8] and implemented [9]. The SR receiver in RF shows better performance than that in baseband [8].

However, in those studies, the transmitted signal was not bandlimited. In a real situation, transmitted signals must be bandlimited by a transmit-filter for removing spurious power and preventing interference with other channels. Filtering causes a fluctuation of the signal amplitude and fluctuating amplitude affects an SR receiver because the SR effect highly depends on the input signal amplitude.

Herein, we evaluate the effect of a filtered signal on the SR Receiver. We consider RF binary phase shift keying (BPSK) signals with and without filtering and evaluate performance using bit-error rate (BER).

This paper is organized as follows. First, we show the system model of an SR receiver with filtering in Sec. 2. In Sec. 3, we evaluate experimentally the BER performance of the SR receiver with and without filtering, and compare those results. Conclusions are given in Sec. 4.

2. System model

Figure 1 shows the system model of the SR receiver. We assume that the transmitted signal is filtered and the received signal power is too weak to be received by a conventional receiver, i.e., the received signal level is lower than the minimum level that the receiver can detect. In such a situation, we use an SR receiver to receive such a weak (subthreshold) signal and communicate with it.

2.1. Transmitter

We assume that a BPSK signal in baseband $b(t)$ is given by

$$b(t) = \sum_i d_i g(t - iT),$$

where $d_i$ is the $i$th symbol of a binary data sequence, which takes $\{\pm 1\}$ and $g(t)$ is a rectangular pulse that has duration $T$ with unity amplitude.
We denote \( b_f(t) \) as the BPSK signal bandlimited by the root-raised-cosine (RRC) filter \( x_{RRC}(t) \), given by

\[
b_f(t) = x_{RRC}(t) * b(t),
\]

where \(*\) represents convolution. The frequency characteristic of \( x_{RRC}(t) \), \( X_{RRC}(f) \) is given as

\[
X_{RRC}(f) = \begin{cases} \sqrt{\frac{2}{\pi}} \frac{\sin\left(\frac{\pi f}{2\sqrt\alpha}\right)}{\pi f} & 0 \leq |f| \leq \frac{1-\alpha}{2\sqrt\alpha} \\ \sqrt{\frac{2}{\pi}} \left[1 + \cos\left(\frac{\alpha}{\pi} \left(|f| - \frac{1-\alpha}{2\sqrt\alpha}\right)\right)\right] & \frac{1-\alpha}{2\sqrt\alpha} \leq |f| \leq \frac{1+\alpha}{2\sqrt\alpha} \\ 0 & |f| > \frac{1+\alpha}{2\sqrt\alpha} \end{cases}
\]

(3)

The RRC filter is commonly used in wireless communication systems, because it helps in minimizing intersymbol interference. The roll-off factor \( \alpha \) determines the excess bandwidth of a signal. When \( \alpha = 0.5 \), the excess bandwidth is 50%. A small roll-off factor results in strict bandlimiting, but causes time-domain ripples and distortion. Therefore, in wireless communication systems, the roll-off factor is generally set to 0.2-0.5.

Figure 2 shows examples of a power spectrum density of a BPSK signal with (a) non-filtering and (b) filtering with an RRC filter. The non-filtered signal illustrated in Fig. 2 (a) has a spurious power spectrum. The filtered signal is bandlimited and the removing spurious power spectrum is as shown in Fig. 2 (b).

After applying the RRC filter, \( b_f(t) \) is upconverted to carrier frequency \( f_c \) and the RF BPSK signal \( s(t) \) is transmitted, which is given by

\[
s(t) = b_f(t) \sin f_c t.
\]

(4)

The transmitted signal \( s(t) \) is propagated in a wireless communication channel. Through that channel, the transmitted signal is attenuated by factor \( \beta \), and a channel noise \( n_c(t) \) is added to the attenuated signal. In general, channel noise \( n_c(t) \) is assumed to be a zero-mean Gaussian noise. The received signal \( r(t) \) is given by

\[
r(t) = \beta s(t) + n_c(t).
\]

(5)

2.2. SR receiver

We construct an SR receiver composed of an SR system and a conventional RF receiver as shown in Fig. 1. The SR system is connected to the front stage a conventional RF receiver.

We assume that the attenuated signal \( \beta s(t) \) is lower than the minimum level \( \xi_{RX} \) that the receiver can detect. Thus

\[
|\beta s(t)| < \xi_{RX}.
\]

(6)

In this system, the received signal \( r(t) \) is the input to the SR system. The SR system is composed of intentional noise \( n_{SR} \) and a nonlinear-device exhibiting SR. We use the SR system to enhance the received signal, and then the conventional receiver can process the output signal of the SR system \( r_{SR}(t) \). When channel noise power is not large enough for the SR system to exhibit its optimal performance, we require additional intentional noise \( n_{SR} \) and adjustment of the noise power to the optimal noise power of the SR [7,9]. Note that the output level of the SR system is enough larger than \( \xi_{RX} \).

If the SR system has a threshold smaller than or equal to the sensitivity of the conventional receiver, the SR receiver can have better performance than that of the conventional one. This is because unlike the conventional receiver, the SR receiver can detect subthreshold signal.

The conventional receiver applies the RRC filter to the received signal in baseband. The receiver-filter is the same as the transmit-filter.

2.3. Filtering

In this section, we discuss how filtering affects the SR receiver.

Figure 3 shows examples of a BPSK signal in the time-domain, with (a) non-filtering and (b) filtering with an RRC filter. In Fig. 3 (a), the signal peak level is constant in its symbol duration. However, in Fig. 3 (b), the peak level fluctuates significantly fluctuate. This fluctuation is caused by filtering and cutting the spurious power spectrum.

We focus on the fluctuation of the signal level caused...
by filtering. In the SR system, subthreshold signals can be detected by adding intentional noise and then the signal peak level exceeds the threshold. The performance of the SR system depends on the difference between the threshold and amplitude. Smaller difference exhibit better SR performance. In this sense, the SR system is sensitive to the received signal level, and the fluctuation by filtering has some effects on its performance.

In this paper, we evaluate the effect of fluctuation by filtering. The SR performances between the filtered signal and non-filtered signals are not identical because the large-amplitude parts of the signal are easily detectable while the small-amplitude parts are hard to detect. The unevenness of the received signal amplitude may have some effects on the SR receiver’s performance.

3. Experiment

In this section, we experimentally evaluate the effect of a filtered BPSK signal on the SR receiver. We use BPSK signals with and without filtering, and evaluate the effect of filtering by comparing the results.

3.1. BER measurement

The BER measurement system is shown in Fig. 4, and the parameter settings are shown in Table 1 and 2. We use a Schmitt trigger as the nonlinear device, and a software-defined radio (SDR) transceiver (NI USRP-2920) as a conventional transceiver. The baseband signals are filtered by digital signal processing. The threshold level of the Schmitt trigger $\xi_{SR}$ is higher than the that of the SDR receiver.

For SR at RF, the Schmitt trigger is designed with a high speed comparator (Analog Devices ADCMP607), which has a wide input bandwidth of 750MHz [10]. The input noise, which is the sum of the channel and intentional noise, is assumed to be a zero-mean Gaussian noise. We add the noise of bandwidth of 100MHz using a the vector signal generator (Agilent Technologies N5182A).

The transmitted signal is attenuated by an attenuator, and the signal with noise is fed into the Schmitt trigger. Table 2 shows the amplitude of the received signals, which are sub-thresholds of the Schmitt trigger. Both of the filtered and non-filtered signals are set to have the same mean square amplitude, meaning they have the same signal power. This parameter setting result in the same BER performance as that of a conventional receiver.

The filter length determines the order of the finite impulse response filter. We set this parameter large enough for the experiment. We set the roll-off factor $\alpha=0.5$, which is the practical value.

Figure 5 shows the BER performance of the SR receiver. As we see from the figure, the BER performance of the SR receiver improves in a specific noise power region. This is a typical phenomenon exhibiting SR. In this region, the input noise power is optimal for SR, so the SR receiver can detect a subthreshold signal and the BER performance can be improved.

3.2. Effect of filtering on BER performance

Figure 6 shows the result in the region where noise power is lower than optimal noise power of the SR. In this region, filtered BPSK shows better BER performance than non-filtered BPSK.

The reason exhibiting such a result is the maximum amplitude. In SR system, signal can be detected by exceeding the threshold. In the region where noise power is lower than optimal noise power of the SR, BER performance are diminished because a weak signal having low power noise cannot exceed the threshold and hence cannot be detected. In the SR system, the suprathreshold signal is critical for detecting a weak signal. As Table 2 shows, the maximum amplitude of the filtered (RRC) BPSK signal is larger than

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Threshold of Schmitt Trigger $\xi_{SR}$</td>
<td>100[mV]</td>
</tr>
<tr>
<td>Modulation scheme</td>
<td>BPSK</td>
</tr>
<tr>
<td>Symbol rate $1/T$</td>
<td>250[kHz]</td>
</tr>
<tr>
<td>Carrier frequency $f_c$</td>
<td>70[MHz]</td>
</tr>
<tr>
<td>Transmitted data bits</td>
<td>10000</td>
</tr>
<tr>
<td>Number of trials</td>
<td>100</td>
</tr>
<tr>
<td>Noise distribution</td>
<td>Gaussian</td>
</tr>
<tr>
<td>Noise bandwidth</td>
<td>100[MHz]</td>
</tr>
<tr>
<td>I/Q sampling rate</td>
<td>4[MHz]</td>
</tr>
<tr>
<td>Filter type</td>
<td>RRC</td>
</tr>
<tr>
<td>Filter length</td>
<td>8</td>
</tr>
<tr>
<td>Roll-off factor $\alpha$</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Table 2: Signal amplitude parameter

<table>
<thead>
<tr>
<th>Filter type</th>
<th>Average amplitude[mV]</th>
<th>Maximum amplitude[mV]</th>
<th>Mean square amplitude[mV$^2$]</th>
</tr>
</thead>
<tbody>
<tr>
<td>none</td>
<td>35</td>
<td>32.78</td>
<td>1.225</td>
</tr>
<tr>
<td>RRC</td>
<td>35</td>
<td>51.1</td>
<td>1.225</td>
</tr>
</tbody>
</table>
4. Conclusion

In this paper, we experimentally evaluated the BER performance of filtered BPSK on an SR receiver. The result shows that filtering improves the BER performance of the SR receiver because amplitude fluctuation contributes to improving the SR effect. This result indicates the applicability of the SR receiver to bandlimited BPSK signals.

Acknowledgments

The authors would like to note that discussions with Prof. Masaaki Katayama, associate Prof. Hiraku Okada and assistant Prof. Kentaro Kobayashi have been illuminating this study. A part of this work is supported by KAKENHI, Grant-in-Aid for Scientific Research 26630174, and THE HORIZ SCIENCES AND ARTS FOUNDATION.

References

Robust Detection of Surface Myoelectric Signal Using A Nonlinear Device Network for Intuitive Man-Machine Interface

Yuki Inden, Kento Shirata, and Seiya Kasai

Research Center for Integrated Quantum Electronics and Graduate School of Information Science & Technology, Hokkaido University
N13, W8, Sapporo 060-8628, Japan
Email: inden@rciqe.hokudai.ac.jp, kasai@rciqe.hokudai.ac.jp

Abstract—Robust myoelectric signal detection using a nonlinear device network and its application to man-machine interface are investigated. The detection system includes the Schmitt trigger network for detecting weak myoelectric signal using stochastic resonance (SR) effect together with multiple surface electrodes made of the carbon nanotube (CNT) composite papers. The system can robustly detect the signal even with extra motion of the body, whereas the conventional system suffers from large noise and cannot distinguish the signal in such case. The robot arm implementing the SR-based detection system is successfully controlled by the gesture of the subject even with his extra motion.

1. Introduction

Myoelectric signal is the active potential generated when the muscle is tensioned. The man-machine interface (MMI) applying the myoelectric signal can provide the intuitive machine control for the users: gesture control. Considering the easy use, the myoelectric signal should be taken from the surface of the body, instead of by the insertion of an electrode into the body. However, the signal generated inside the body is attenuated and is easily buried in noise, although the conventional detection technique successfully detects the weak myoelectric signal by differential amplification, it inevitably loses the function when the contact between the surface electrode and body is fluctuated. The concept of our technique is to detect the weak myoelectric signal using noise and fluctuation through stochastic resonance (SR), in which the response to the weak signal is optimized or enhanced by adding noise [1,2]. We demonstrate the robust myoelectric signal detection by the Schmitt trigger network causing the SR and its feasibility for the intuitive MMI through the robot arm control. So far, in the biological signal sensing research field, the SR was only investigated to enhance the sensitivity of the subject by adding noise to himself [5]. Recently our group achieved the high sensitive and robust myoelectric signal detection using the SR [3,4]. The contribution of this paper is the demonstration of the feasibility of our SR-based myoelectric signal detection system for the muscle tension detection, motion discrimination, and robot arm control.

2. Myoelectric Signal Detection System

Figure 1 shows our SR-based myoelectric signal detection system. The system integrates eight Schmitt triggers as nonlinear devices with hysteresis to cause the stochastic resonance. The devices form a summing network to obtain enhanced response in accordance with the framework of Collins’s system [6]. The previous stage of the Schmitt trigger includes high pass filters (HPF) for offset canceling, a preamplifier, and a band elimination filter (BEF) for filtering 50 Hz hum noise. The second HPF prior to the Schmitt trigger is inserted to completely remove the offset fluctuation. Thus the imposed noise to the signal almost passes to the input of each Schmitt trigger. Each input of the first stage of the system is connected to a carbon nanotube composite paper (CNTcp)-based surface electrode. The performance of the CNTcp electrode is comparable to the conventional AgCl/Ag electrode, even with relatively high sheet resistivity [4]. The CNTcp electrode has advantages in the ease of processing, flexibility, soft texture, and disposability. The SR-based system essentially achieves the high signal-to-noise ratio (SNR) by the threshold transfer characteristic of the Schmitt trigger, similar to the high

Fig. 1 SR-based myoelectric signal detection system.
SNR in the digital signal transmission. This signal truncation process is effective, because the intensity of the muscle tension is basically coded into the density of the action potential pulse train, not the amplitude of the pulse. On the other hand, the surface myoelectric signal is composed of the action potentials from many muscle fibers and the potential generated deep in the body is attenuated remarkably. The SR-based system detects such weakened signal using the stochastic resonance caused in the Schmitt trigger. The multiple surface electrodes average out the contact fluctuation that occurs in uncorrelated manner between the electrodes. We already confirmed that the SR can be caused on the aperiodic myoelectric signal in the nonlinear device [3].

3. Detection Characteristics

Figure 2 shows the measured surface myoelectric signals using our technique and conventional bipolar lead technique. The signals were taken from the surface of the forearm of the subject without and with the movement of the shoulder as extra motion. The hysteresis width of each Schmitt trigger device was adjusted to be approximately 100 mV. Without the shoulder motion, both techniques detect the myoelectric signal clearly. The output waveform of the SR-based system showed the pulses having mostly uniform height. However each pulse had different width: the amplitude of the original myoelectric signal was reflected in the width of the output pulse. In the case with the motion of the shoulder, the output from the conventional technique was tremendously disordered all the time and it was impossible to distinguish the myoelectric signal from the noisy waveform. The large noise in the bipolar lead was generated because the balance of the contacts of the two surface electrodes was broken and the fluctuation is not canceled out but amplified. On the other hand, our technique could successfully detect the myoelectric signal even with the motion of the shoulder. It should be noted that the noises generated from the contact fluctuation were uncorrelated between the electrodes and this provided a positive effect on the detection performance in terms of the stochastic resonance in the summing network [6].

The observed waveforms show that high sensitivity and noise robustness of our system is attributed to the combination of the SR mechanism and the noise rejection by the double thresholds. Myoelectric signal is represented by the pulse train. Its bandwidth is 2 Hz ~ 10 kHz and widely overlaps with that of the noise. Considering these points, rejection of the signal component out of the thresholds is a rational way compared to the filtering in the frequency domain. On the other hand, the weak myoelectric signal generated in the deep inside of the body will be filtered out when the noise is eliminated using a low pass filter (LPF). The SR mechanism helps to detect the such signal component.

Table 1 summarizes the evaluated output SNRs for the various detection techniques. The unipolar lead detects the signal using one surface electrode and filters the noise using a low pass filter (LPF) with a linear amplifier. The performance of the commercially available device is also shown. The SR-based technique showed the highest SNR in the examined ones in both without and with the motion of the body. The myoelectric signal detector in the recent commercial myoelectric prosthesis has a very powerful dynamic filter that can detect the signal even in the motion of body. However it needs the learning process and takes much machine power for signal processing. Our technique can reduce such machine cost and is expected to give faster response. In addition, our technique can achieve high SNR comparable to that using the needle electrode. The surface electrode technique detects the signal from a bundle of the muscle fibers, whereas the needle electrode can detect the action potentials from a few muscle fibers. The former is appropriate for the MMI application and the latter is necessary for medical examination and analysis.

<table>
<thead>
<tr>
<th></th>
<th>Stationary</th>
<th>In Motion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unipolar lead</td>
<td>15 dB</td>
<td>10 dB</td>
</tr>
<tr>
<td>Bipolar lead (Conventional)</td>
<td>20 dB</td>
<td>undetectable</td>
</tr>
<tr>
<td>Commercial detector</td>
<td>16 dB</td>
<td>undetectable</td>
</tr>
<tr>
<td>This work</td>
<td>40 dB</td>
<td>20 dB</td>
</tr>
</tbody>
</table>

Table 1 Evaluated signal-to-noise ratio (SNR) for various surface myoelectric signal detection techniques.
Figure 3 shows the forearm tension dependence of the detected myoelectric signals. The tension was quantitatively measured using a hand dynamometer. The density of the action potentials obtained by the conventional technique was clearly changed depending on the strength of the tension. The output waveforms in the SR-based system also depended on the tension, however, the change was not so obvious as in the conventional technique when the tension was 30% and 50%. We evaluated the power of the obtained signal, $P_S$. In the case of the conventional technique, $P_S$ for 30%, 50% and 70% was 1.36 V$^2$, 2.16 V$^2$, and 7.17 V$^2$, respectively. On the other hand, in the case of the SR-based technique, $P_S$ for 30%, 50% and 70% was 4.70 V$^2$, 5.77 V$^2$, and 6.35 V$^2$, respectively. These results showed that $P_S$ in the SR-based technique gave linear response compared to the conventional technique. The SR-based technique has possibility to detect weak tension better than the conventional technique.

4. Motion Identification

For the MMI by gesture control, it is necessary to discriminate and identify the various motions. Physiology suggests that a person has various muscles and each motion of the subject is attributed to the different muscle. Therefore the identification is achieved by analyzing the myoelectric signals taken from several surface electrodes on the appropriate positions of the body. We examined the identification of the wrist motion, palmar and dorsal flexion, using our detection system. Mainly two different muscles contribute to the two motions of the wrist. Then the two multiple surface electrode arrays, electrodes A and B, were attached on the forearm near the related muscle positions as shown in Fig. 4 and the signals were taken using the two detectors independently. Measured waveforms are shown in Fig. 5. When the palmar flexion took place, the myoelectric signal was induced only in the electrode A. On the other hand the signal was induced in the electrode B when the dorsal flexion took place. Then the wrist movement was identified by evaluating the difference of the two myoelectric signal power. We found that the SR-based system gave clear power difference between the two wrist motion compared to that of the conventional bipolar detection system.

5. Robot Arm Control

To demonstrate the feasibility of the SR-based myoelectric signal detection technique for the robust MMI, we designed the robot arm control system implementing the SR-based detection technique as shown in Fig. 6(a). This system consisted of two detectors together with a microcomputer. The system identified the two motions of the wrist, palmar and dorsal flexion. In accordance with the identified motion of the subject as described in the previous section, the wrist of the robot arm moved upward or downward. The CNTcp-based surface electrode array was easily attached to the arm of the subject using an arm band, without electrolytic paste and tight binding. A snapshot of the demonstration is shown in Fig. 6(b) (the
movie of this experiment will be shown at the presentation). The robot arm could be correctly controlled in accordance with the motion of the wrist of the subject. In addition, such controllability was maintained even with the extra motion of the shoulder, whereas the system using the conventional technique became uncontrollable in such case. The obtained results demonstrated the feasibility of our technique for the robust MMI.

6. Conclusions

Robust detection of the myoelectric signal using a 8 Schmitt trigger summing network and its application to man-machine interface were presented. Weak myoelectric signal was detected using the stochastic resonance in the Schmitt trigger network together with multiple surface electrodes made of the carbon nanotube (CNT) composite papers. The SR-based system could robustly detect the signal even with extra motion of the body, although the conventional system missed the signal in such case. The feasibility of the SR-based myoelectric signal detection system for the intuitive MMI was demonstrated by the robust control of the robot arm implementing the SR-based system.

Acknowledgments

The authors thank Mr. H. Nakamura, Mr. Y. Hagiwara, and Mr. S. Kaeriyama of STARC for their valuable discussions and supports. They also thank Prof. T. Oya of Yokohama National University and KJ Specialty Paper Co. Ltd. for providing CNT-based materials. This work is partly supported by STARC program and a Grant-in-Aid for Scientific Research on Innovative Areas “Molecular Architectonics” (#25110001 and #25110013) from the MEXT, Japan.

References

Subcarrier Allocation in Multi-carrier DCSK System for Performance Enhancement

Hua Yang† and Guo-ping Jiang‡

†Faculty of College of Electronic Science and Engineering, Nanjing University of Posts and Telecommunications
9 WenYuan Road, Nanjing 210023, China
‡Faculty of College of Automation, Nanjing University of Posts and Telecommunications
9 WenYuan Road, Nanjing 210023, China
Email: yangh@njupt.edu.cn, jianggp@njupt.edu.cn

Abstract—Recently, multicarrier differential chaos shift keying (MC-DCSK) modulation has been proposed to avoid the usage of delay lines as well as to obtain higher data rate in comparison to DSK. Thanks to better energy efficiency, MC-DCSK outperforms DCSK in bit error rate (BER) performance, but it still performs much worse than conventional coherent BPSK. This paper shows how to get great improvement in BER performance of MC-DCSK by assigning multiple subcarriers to the reference signal. A novel subcarrier allocated MC-DCSK system is proposed, where every reference signal is transmitted for multiple times on different subcarriers and all received copies are averaged for noise cleaning. With the optimal subcarrier allocation strategy, the proposed system can achieve the lowest BER that is almost identical to that of the conventional coherent BPSK system.

1. Introduction
In recent years, considerable attention has been paid to designing chaos-based digital modulations [1]-[4]. In these systems, spectrum spreading and digital modulation are performed simultaneously by mapping data symbols to dissimilar wideband chaotic signals. Since chaotic signals serving as the carriers are non-periodic, difficult to predict and quite easy to generate, chaos-based digital modulation schemes not only enjoy all the merits of traditional spread spectrum (SS) systems (i.e., low probability of detection, anti-jamming, mitigation of multi-path fading and so on) but also show good communication security with low cost [5].

Up to now, many schemes have been proposed [6]-[9], in which differential chaos shift keying (DCSK) and its frequency modulated version have attracted more research interests for good performance and low cost. However, DCSK suffers from low bit rate and delay line problem in UWB communications, as the reference and data-bearing signals are transmitted in sequential time periods [10]. For high bit rate as well as delay line removal, multi-carrier DCSK (MC-DCSK) has been recently proposed in [11], where one reference signal together with multiple data-bearing signals are sent simultaneously on all subcarriers. Even though MC-DCSK could outperform DCSK in noise performance due to higher energy efficiency, the noise performance of MC-DCSK is still much worse than that of the conventional coherent BPSK.

To remarkably improve the noise performance of MC-DCSK, subcarrier allocation strategy is considered here and a novel MC-DCSK system using multiple subcarriers to send repeated reference signals is proposed in this paper. Our goal is to exploit this repetition of received reference signals for noise reduction, which can be achieved by averaging all corrupted copies of any reference signal. The bit error rate (BER) performance of the proposed system is evaluated over additive white Gaussian noise (AWGN) channel by simulation, and the optimal BER performance is obtained with the optimal subcarrier allocation strategy. In addition, to clean the noisy data-bearing signals, noise reduction algorithm given in [12] is also applied to the proposed system. Relevant performance comparisons are given, which confirms the significant advantages of the proposed subcarrier allocation strategy.

2. Subcarrier allocated MC-DCSK system
In the original MC-DCSK system in [11], a serial high rate bit stream is converted into multiple parallel low rate bit sub-streams. Data bits in all sub-streams use the same chaotic message bearer, which will be sent along with all bit-streams in a parallel way on different subcarriers.

Like the original MC-DCSK, subcarrier allocated MC-DCSK (SA-MC-DCSK) also convert a serial bit stream into multiple sub-streams, all of which use a chaotic signal $x(t)$ as the message bearer, denoted as

$$x(t)=\sum_{k=1}^{B_s} x_k(t-k T_s)$$ (1)

where $\beta$ is the spreading factor, $x_k$ is the $k$-th sample of chaotic sequence, and $h_r(t)$ is the impulse response of a square-root-raised cosine filter with a roll-off factor $\alpha$, normalized energy and duration $T_s$. To satisfy the Nyquist criterion, frequency spectrum of $h_r(t)$ is limited to $[-B_c/2, B_c/2]$ with $B_c=(1+\alpha)/T_s$.

Unlike the original MC-DCSK, several copies, rather than a single copy, of the message bearer in (1) (i.e., the reference signal) accompanied by all bit-streams are sent. This could be achieved by allocating more subcarriers to the reference signal.
Data

\[ r_j(t = kT) = \begin{cases} x_j + n_{j,k}, & 1 \leq j \leq N \\ b_{j,N}x_j + n_{j,k}, & N < j \leq M \end{cases} \]

(3)

Here, we assume that the received signal is only corrupted by an AWGN noise with PSD of \( N_0/2 \). \( n_{j,k} \) is the \( k \)-th sample of the noise that corrupts the signal transmitted on the \( j \)-th subcarrier.

The reference signal can be estimated by averaging the outputs of the upper \( N \) samplers, which can be denoted as

\[ y_j = x_j + \frac{1}{N} \sum_{k=1}^{N} n_{j,k} \]

(4)

Since noises that pollute the signals sent on different subcarriers are independent and identically distributed, the variance of the noise content in (4) is \( N \) times smaller than that of \( n_{j,k} \), contributing positively to the improvement in BER performance.

In Fig.2, data-bearing signal that carries the bit of the \( i \)-th data stream is recovered as the output of the \( i+N \)-th sampler

\[ d_{i,k} = b_i x_k + n_{i,k} x_k, \quad i = 1, \ldots, M - N \]

(5)

By correlating the estimated reference signal in (4) with recovered data-bearing signal in (5), the decision variable for the bit of the \( i \)-th data stream is

\[ Z_i = \sum_{k=1}^{N} y_k \cdot d_{i,k}, \quad i = 1, \ldots, M - N \]

(6)

According to [11], energy efficiency of SA-MC-DCSK can be evaluated by the Data-energy-to-Bit-energy Ratio (DBR) represented as

\[ \text{DBR} = \frac{M - N}{M} \]

(7)

The DBR of the original MC-DCSK system is equal to that of SA-MC-DCSK with \( N = 1 \) as the proposed system turns into the original MC-DCSK system when \( N = 1 \).

For comparison, the DBRs of subcarrier allocated MC-DCSK with various \( N \) are plotted in Fig.3 against the number of subcarriers \( M \). Clearly, SA-MC-DCSK with \( N = 1 \) (i.e., the original MC-DCSK) has the highest energy efficiency as merely one subcarrier is dedicated to sending the reference. With more subcarriers being allocated to the reference signal, energy efficiency declines accordingly. This DBR decrease will contribute negatively to the BER performance. However, for larger \( N \), more received copies of any reference signal are averaged, which contributes positively to the BER performance as a result of weaker noises in decision variables. Therefore, we believe that the proposed system shows optimal BER performance when a balance between these two contributions is achieved. In
the following section, we will discuss how to get the optimal BER performance.

![Fig. 3: DBRs of subcarrier allocated MC-DCSK](image)

**4. Performance evaluation**

In this section, the proposed system as well as the original MC-DCSK is simulated over AWGN channel. In these two systems, chaotic sequences are generated by the logistic map $x_{i+1} = 1 - 2x_i^2$ in [5]. The roll-off factor $\alpha$ is set to 0.25. The total bandwidth of all subcarriers is 4MHz. For fixed bit duration $T$ and subcarrier number $M$, the spreading factor can be computed by $\beta = T B / M (1 + \alpha)$.

To study the optimal performance behavior of SA-MC-DCSK, simulated BERs of the proposed scheme are given in Fig.4 for various subcarrier numbers $M$. All curves are plotted against number $N$ of subcarriers allocated to the reference signal under a certain signal-to-noise ratio (SNR) level.

![Fig. 4: Relationship between $N$ and the BERs of subcarrier allocated MC-DCSK with $\beta=128$ and $E_b/N_0 = 10$dB](image)

When $N$ increases, it is observed in Fig.4 that the BERs under a fixed SNR level first drop and then tend to rise. This interesting phenomenon is caused by the interaction between reduced noise in the estimated reference in (4) and decreased DBR in (7). On one hand, the estimated reference in (4) becomes cleaner if $N$ grows, making BER performance improved. On the other hand, with more subcarriers being occupied by the reference signals, DBR in (7) becomes smaller, leading to degraded performance. As a result, the lowest BER can be obtained by SA-MC-DCSK if an optimal $N$ is used. For example, it is observed in Fig.4 that with $\beta=128$ and $E_b/N_0 = 10$dB, the optimal values of $N$ are 22, 14, 9 and 4 for $M=128$, 64, 32 and 16 respectively.

Besides, it is also noticed in Fig.4 that the distance between the lowest BER and the BER with $N=1$ (i.e., BER of the original MC-DCSK) become larger if $M$ increases. This means that the performance improvement brought by applying subcarrier allocation grows with the total number of subcarriers.

![Fig. 5: BER comparison between MC-DCSK and SA-MC-DCSK with the optimal $N$](image)

To further investigate the performance improvement achieved by subcarrier allocation, simulated BER curves of SA-MC-DCSK with the optimal $N$ (labeled as ‘Optimal SA-MC-DCSK’) are plotted in Fig.5. Here, BER curves of the original MC-DCSK and conventional coherent BPSK are also given for comparison. It is obvious that the proposed system with optimal $N$ performs much better than the original MC-DCSK. This superiority in performance grows with spreading factor $\beta$. For instance, the obtained BER gain is 3-4dB when $\beta$ is 16, while 4-5dB gain is achieved when $\beta$ equals to 128. By assigning the optimal number of subcarriers to the reference signal in MC-DCSK, the distance between the performances of MC-DCSK and the conventional coherent BPSK has been shortened to less than 1dB.

The noise performance of the proposed system can be further improved if the noisy data-bearing signals are also cleaned. Considering the fact that each data-bearing signal, either in its identical or inverted version, is transmitted for multiple times in the proposed system, all received data-bearing signals could also be averaged for noise reduction if the modulations are removed. Fortunately, this problem is perfectly solved by a simple noise reduction algorithm proposed in [12]. For this reason, the algorithm in [12] is employed here to reduce the noise contents in all received data-bearing signals. This could be achieved by replacing the data-bearing signal $d_{i,k}$ in decision variable in (6) with $\tilde{d}_{i,k}$, denoted as
Fig. 6 The optimal BER performance of SA-MC-DCSK with noise reduction algorithm in [12].

Fig. 6 evaluates the effect of this simple algorithm on BER performance of SA-MC-DCSK. In this figure, the spreading factor is 16. It is interesting to find that, with the help of this noise reduction algorithm, the optimal BER performance obtained by subcarrier allocated MC-DCSK (labeled as ‘Optimal SA-MC-DCSK with noise reduction’) happens to be almost same to that of the conventional coherent BPSK.

4. Conclusion

In this paper, a novel MC-DCSK system is proposed based on subcarrier allocation. In this new system, more subcarriers are dedicated to sending repeated copies of the reference signal, which is shared by all data-bearing signals transmitted on the remaining subcarriers. Before performing correlation at the receiver side, all received copies of each reference signal are averaged for noise reduction.

With the optimal number of subcarriers allocated to the reference signal, the proposed system achieves the best BER performance which is 1dB worse than that of the conventional BPSK system.

Combined with the algorithm that reduces the noises in all received data-bearing signals, the proposed system can get the optimal BER performance almost same to that of the conventional coherent BPSK system.

Acknowledgments

This work has been supported in part by the National Natural Science Foundation of China under grants of 61374180 and 61401226, and the Natural Science Foundation of Jiangsu Province under grant of 16KJB510045.

References

Robust Scale-free Luby Transform Code and Its Performance

Yuli Zhao and Francis C.M. Lau, Zhiliang Zhu, Hai Yu and Li Guo

‡Software College, Northeastern University, Shenyang, China
#School of Computer Science and Engineering, Northeastern University, Shenyang, China
‡Department of Electronic & Information Engineering, Hong Kong Polytechnic University, Hong Kong;
and Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China.
Email: zhaoyl@swc.neu.edu.cn, francis-cm.lau@polyu.edu.hk

Abstract—Compared with Luby Transform (LT) codes having an ideal/robust soliton degree distribution, LT codes with encoded-symbol degree following a modified power-law distribution (scale-free LT codes) have been shown to possess a higher probability of successful decoding and a lower encoding/decoding complexity when the information symbol length ranges from 512 to 2048. In an attempt to reduce the size of the initial ripple set of scale-free LT (SF-LT) codes so as to prevent the ripple set from becoming empty, a new class of LT codes, namely robust SF-LT (RSF-LT) codes, is proposed in this work. The performance and characteristics of the proposed RSF-LT code are compared with those of other LT codes including robust LT code, SF-LT code, and “LT code with decreasing ripple set”. Results show that the proposed RSF-LT code outperforms the other codes with respect to average overhead, encoding/decoding efficiency and probability of successful decoding.

1. Introduction

Luby transform (LT) codes are the first type of practical rateless code [1]. It is originally designed for reliable data transmission over a binary erasure channel (BEC), which is suitable for the modeling of the Internet. Recently, LT codes have also been discussed for use in mobile multimedia broadcasting, wireless sensor networks, etc. [2, 3].

An LT code is capable of generating an unlimited number of encoded symbols based on a source message of length $K$. Regardless of the erasure probability of a BEC, an LT decoder can recover the original $K$ input symbols when $(1+\alpha)\times K$ encoded symbols have been received. Here $\alpha$ is a real number slightly larger than 0. Consequently, LT code is a near-optimal channel code for all erasure channels.

Whether an LT code is well designed or not is determined by the degree distribution of its encoded symbols [4]. The ideal soliton distribution is the first degree distribution used to construct LT codes [1]. LT codes based on such a distribution can theoretically keep the ripple size always equaling one in the decoding process. Therefore, such a design avoids any redundancy and is optimal. However, any fluctuation around this expected behavior results in a lack of degree-1 encoded symbols and hence an un-

successful decoding [5]. To deal with the above problem, a robust soliton distribution has further been proposed and it aims at maintaining a ripple size larger than one in the whole decoding process [1]. Results show that LT codes based on the robust soliton distribution outperform the original LT codes.

In [6], an LT code that can maintain the ripple size to a pre-defined constant during the decoding process has been proposed. In [7], LT codes with decreasing ripple size are designed and analyzed. The results indicate that such LT codes are capable of producing a higher performance. In [8], using the shortest-average-path-length property of scale-free networks, a class of scale-free LT (SF-LT) codes has been proposed. It has further been shown that SF-LT codes outperform LT codes based on robust soliton distribution and LT codes based on suboptimal distribution. In this paper, a new class of LT codes, namely robust SF-LT (RSF-LT) codes, is proposed and investigated.

2. Proposed Robust SF-LT code

In [8], a SF-LT code with the degree of the encoded symbols following a modified power-law distribution has been proposed. Specifically, the distribution is given by

$$\tau(d) = \begin{cases} P_1, & d = 1 \\ Ad^{-\gamma}, & d = 2, 3, ..., K - 1, K. \end{cases}$$

(1)

where $P_1$ is the fraction of degree-1 encoded symbols; $\gamma$ is the characteristic exponent; and $A$ is a normalizing coefficient to ensure $\sum_{d=1}^{K} \tau(d) = 1$.

We defined a released encoded symbol as an encoded symbol whose degree becomes 1 during the iterative decoding process, and a ripple set as the set of input symbols which are connected to the released encoded symbols. Assume that at the end of each iteration in the decoding process, the neighboring input symbols of a newly released encoded symbol are not elements in the ripple set. Suppose $(1+\alpha)\times K$ encoded symbols have been received to recover $K$ input symbols. Then, the theoretical evolution of
where $\rho(d)$ is the ideal soliton distribution expressed as

$$\rho(d) = \begin{cases} 
\frac{1}{K} & d = 1, \\
\frac{1}{d(d-1)} & d = 2, 3, \ldots, K
\end{cases}$$

and $\tau(d)$ is the modified power-law degree distribution given in Eq.(1).

The encoding and decoding complexity are both going to scale linearly with the number of edges in the Tanner graph. It has been shown in [5] that when the number of encoded symbols received is close to Shannon’s optimal, i.e. $K$ encoded symbols, the average degree of each encoded symbol should be at least $\ln K$ for the sake of making the decoding possible. Consequently, it is necessary to ensure that $\sum_{i=1}^{K} \{d\rho(d)\} > \ln K$ when the parameters $P_1, \gamma$ and $A$ are selected for the RSF-LT codes.

3. Results and Discussions

In this section, the characteristics and performance of (i) LT codes based on robust soliton degree distribution [1]; (ii) SF-LT codes [8]; (iii) LT code with decreasing ripple size in [7] and (iv) proposed RSF-LT codes are compared. The particular LT codes to be studied are as follows.

- Robust LT code: LT codes based on robust soliton distribution with parameters $\beta = 0.1$ and $\delta = 1$ have been proven to provide the smallest average overhead factor [10]. Such robust LT codes are used here.
- SF-LT1: SF-LT code using $P_1 = 0.1$ and $\gamma = 2.0$
- SF-LT2: SF-LT code using $P_1 = 0.09$ and $\gamma = 2.1$
- RSF-LT1: RSF-LT code using $P_1 = 0.1$ and $\gamma = 1.9$
- RSF-LT2: RSF-LT code using $P_1 = 0.1$ and $\gamma = 2.1$
- RSF-LT3: RSF-LT code using $P_1 = 0.1$ and $\gamma = 2.0$
- RSF-LT4: RSF-LT code using $P_1 = 0.09$ and $\gamma = 2.1$
- LT code in [7]: LT code with decreasing ripple set proposed in [7] using $n = 1075, R = 21$ when $K = 1024$; and $n = 2108, R = 25$ when $K = 2048$.

3.1. Theoretical Evolution of the Ripple size

The theoretical evolution of the ripple size for the proposed RSF-LT code is evaluated by substituting Eq.(3) into Eq.(2). The results are plotted, together with those of robust LT code, SF-LT codes and LT code in [7], in Fig. 2 when the number of input symbols $K = 1024$. It can be observed that except for the LT code in [7] which has a decreasing ripple size, the ripple sizes of other LT codes decrease initially, and then increase before decrease again.

In this paper, the characteristics of the ideal soliton distribution are applied to the design of SF-LT codes, forming the proposed robust SF-LT code. The aim is to decrease the probability that the ripple set becomes empty.

Definition 1 A robust scale-free LT code, denoted as RSF-LT code for short, is defined as an LT code with the degree of the encoded symbols following a distribution given by

$$\mu(d) = \frac{\rho(d) + \tau(d)}{\sum_{i=1}^{K} (\rho(i) + \tau(i))}$$

Figure 1: The theoretical ripple evolution of the SF-LT codes when $K = 1024$. The ripple size can be calculated using [9]

$$\|\Omega^k(i) = (1 + \alpha)Kp(i) \quad i = 1, 2, \ldots, K$$

$$\|\Omega^L(i) = 1 + (1 - \Omega^L(i))\Omega^L(i+1)$$

$$\|\Omega^L(i) = \Omega^L(i) - \frac{1}{L-1}\Omega^L(i+1)$$

where $\Omega^L(i)$ is the number of degree-$i$ input symbols left in the decoding process when $L$ input symbols remain unprocessed, and $p(i)$ denotes the probability of an encoded symbol having a degree $i$. Theoretically, the ripple size should be no less than 1 throughout the whole decoding process to prevent the decoding process stopped prematurely.

The theoretical evolution of the ripple size for a SF-LT code can be evaluated by substituting Eq.(1) into Eq.(2). Fig. 1 plots the ripple evolution of SF-LT codes when $K = 1024$. The results indicate that the SF-LT code can recover the original input symbols when the overhead factor is $\alpha = 0.06$ using the parameter set $P_1 = 0.09$ and $\gamma = 2.1$. However, when $\alpha$ is reduced to 0.05, the ripple size will become smaller than 1 at a certain point and the SF-LT codes will not be able to recover the input symbols.

In this paper, the characteristics of the ideal soliton distribution are applied to the design of SF-LT codes, forming the proposed robust SF-LT code. The aim is to decrease the probability that the ripple set becomes empty.

### 3.1. Theoretical Evolution of the Ripple size

The theoretical evolution of the ripple size for the proposed RSF-LT code is evaluated by substituting Eq.(3) into Eq.(2). The results are plotted, together with those of robust LT code, SF-LT codes and LT code in [7], in Fig. 2 when the number of input symbols $K = 1024$. It can be observed that except for the LT code in [7] which has a decreasing ripple size, the ripple sizes of other LT codes decrease initially, and then increase before decrease again.

The results in Fig. 2 also indicate that RSF-LT code and RSF-LT code can recover the 1024 input symbols when $(1 + 0.05) \times 1024 \approx 1075$ encoded symbols (i.e., overhead $\alpha = 0.05$) have been received. Furthermore, RSF-LT code can recover all input symbols when the number of received encoded symbols approaches $N = (1 + 0.03) \times 1024 \approx 1055$. For the robust LT code, the ripple set will become smaller than 1 during the evolution if the overhead factor equals $\alpha = 0.06$, implying that the decoding process will fail.
With the same overhead $\alpha = 0.06$, SF-LT2 code can recover all encoded symbols. However, as shown in Fig. 1, SF-LT2 code cannot recover the encoded symbols if the overhead is reduced to $\alpha = 0.05$. For LT codes in [7], the ripple set will not become empty when overhead $\alpha \geq 0.03$. In summary, the results indicate that theoretically the proposed RSF-LT codes can achieve similar or even smaller overhead factor compared with other types of LT codes.

Note that in the iterative decoding process, not all the input symbols connected to a newly released degree-1 encoded symbol will be useful. The reason is that they may already exist in the ripple set. Thus, the actual number of encoded symbol $(1 + \alpha)K$ required to recover $K$ input symbols may be larger than the results revealed in this section. Yet, the theoretical ripple evolution can provide an effective method for selecting parameter sets for RSF-LT codes.

### 3.2. Code Characteristics

The characteristics of the LT codes described in the previous section is further studied over a perfect channel. For each of the LT codes, $M = 2000$ different sets are constructed and evaluated when $K = 1024$ and $K = 2048$. The following symbols are defined.

- $\tilde{d}$: average degree of the encoded symbols
- $\tilde{x}$: average number of XOR operations for generating an encoded symbol
- $\tilde{\phi}$: average number of XOR operations for decoding an LT code over a perfect channel
- $\tilde{\alpha}$: average overhead factor.

Table 1 lists the characteristics of the LT codes under study. It can be observed that when $K = 1024$ and $K = 2048$, all the proposed RSF-LT codes outperform the robust LT code in terms of average overhead factor $\tilde{\alpha}$ and average number of encoding/decoding operations ($\tilde{d}$ and $\tilde{\phi}$). The proposed RSF-LT codes also (i) outperform the SF-LT codes in terms of average overhead factor and achieve similar range of average number of encoding/decoding operations compared with the SF-LT codes; and (ii) outperform the LT code in [7] in terms of the average number of encoding/decoding operations and achieve similar range of average overhead factor compared with LT code in [7].

### 3.3. Decoding Performance over a BEC

The decoding performance of the LT codes over a BEC is simulated with an erasure probability of $P_{era} = 0.1$. Fig. 3 and Fig. 4 plot the probability of successful decoding of the LT codes when $K = 1024$ and $K = 2048$, respectively.

In Fig. 3 and Fig. 4, it can be observed that all the proposed RSF-LT codes and the SF-LT2 code outperform robust LT code when the number of encoded symbols received is relatively small (less than 1220 at $K = 1024$, 2300 at $K = 2048$). When the number of encoded symbols received become large, robust LT code begins to outperform other codes but is still outperformed by RSF-LT1 code and RSF-LT3 code. Compared with LT code in [7], RSF-LT3 code achieves a similar probability of successful decoding when the number of encoded symbols received is relatively small. As the number of encoded symbols received becomes larger, RSF-LT1 code and RSF-LT3 code outperform LT code in [7]. Moreover, RSF-LT2 code and RSF-LT4 code achieve a similar probability of successful decoding with LT code in [7] with the same number of encoded symbols received. Based on the results listed in Table I, it can be further concluded that both RSF-LT1 and RSF-LT3 codes can achieve the best performance in terms of encoding/decoding complexity ($\tilde{d}$ and $\tilde{\phi}$) as well as probability of successful decoding when $K = 1024$ and $K = 2048$. 

![Figure 2: Theoretical ripple evolution of the Robust SF-LT codes, SF-LT codes and Robust LT codes when $K = 1024$.](image-url)

![Figure 3: Probability of successful decoding versus the number of encoded symbols received.](image-url)
Table 1: Code characteristics

<table>
<thead>
<tr>
<th>$K$</th>
<th>Code</th>
<th>$\bar{d}$</th>
<th>$\bar{x}$</th>
<th>$\bar{\phi}$</th>
<th>$\bar{\eta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1024</td>
<td>Robust LT code</td>
<td>9.94</td>
<td>8.94</td>
<td>10722</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>RSF-LT1 code</td>
<td>9.57</td>
<td>8.59</td>
<td>9935</td>
<td>0.115</td>
</tr>
<tr>
<td></td>
<td>RSF-LT2 code</td>
<td>7.49</td>
<td>6.49</td>
<td>7562</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>RSF-LT3 code</td>
<td>8.35</td>
<td>7.34</td>
<td>8241</td>
<td>0.107</td>
</tr>
<tr>
<td>1024</td>
<td>RSF-LT4 code</td>
<td>7.53</td>
<td>6.50</td>
<td>7576</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>SF-LT1 code</td>
<td>9.20</td>
<td>8.19</td>
<td>9655</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td>SF-LT2 code</td>
<td>7.54</td>
<td>6.54</td>
<td>7671</td>
<td>0.125</td>
</tr>
<tr>
<td></td>
<td>LT code in [7]</td>
<td>15.24</td>
<td>14.20</td>
<td>16346</td>
<td>0.118</td>
</tr>
<tr>
<td>2048</td>
<td>Robust LT code</td>
<td>11.15</td>
<td>10.15</td>
<td>23729</td>
<td>0.129</td>
</tr>
<tr>
<td></td>
<td>RSF-LT1 code</td>
<td>10.77</td>
<td>9.74</td>
<td>22067</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td>RSF-LT2 code</td>
<td>8.11</td>
<td>7.09</td>
<td>16339</td>
<td>0.110</td>
</tr>
<tr>
<td></td>
<td>RSF-LT3 code</td>
<td>9.18</td>
<td>8.12</td>
<td>18155</td>
<td>0.085</td>
</tr>
<tr>
<td>2048</td>
<td>RSF-LT4 code</td>
<td>8.14</td>
<td>7.14</td>
<td>16266</td>
<td>0.099</td>
</tr>
<tr>
<td></td>
<td>SF-LT1 code</td>
<td>10.16</td>
<td>9.14</td>
<td>21232</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td>SF-LT2 code</td>
<td>8.08</td>
<td>7.07</td>
<td>16374</td>
<td>0.117</td>
</tr>
<tr>
<td></td>
<td>LT code in [7]</td>
<td>17.51</td>
<td>16.61</td>
<td>37689</td>
<td>0.085</td>
</tr>
</tbody>
</table>

Figure 4: Probability of successful decoding versus the number of encoded symbols received. $K = 2048$ and $P_{\text{err}} = 0.1$.

4. Conclusion

In this paper, a new type of LT code called robust scale-free LT (RSF-LT) code has been proposed. It integrates the characteristics of (i) LT codes based on ideal soliton distribution and (ii) scale-free LT codes. Theoretical analyses on the ripple evolution process indicate that the proposed RSF-LT code outperforms robust LT code, SF-LT code and LT code in [7]. Simulations over a BEC further reveals that among all the LT codes under study, RSF-LT1 code and RSF-LT3 code achieve the best probability of successful decoding as well as the lowest encoding/decoding complexity.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant Nos. 61372095, 61374178, 61402092). The online education research fund of MOE research center for online education, China (Qtone education, Grant No.2016ZD306) and the Ph.D. Start-up Foundation of Liaoning Province, China (Grant No. 201501141).

References


Regional Activation based on P2P Network Architecture

Yutaka Naito†, Chiaki Katsuki*, Nozomi Suehiro* and Takashige Hoshiai*

†Kumamoto City Government
1-1 Tetorihocho, Chuoku, Kumamoto City 860-8601, Japan
*Sojo University
4-22-1 Ikeda, Nishiku, Kumamoto City 860-0082, Japan

Abstract - In this paper, we propose the usage of a P2P network technology for realizing regional activation on regional communities. We describe the network structure about both the inside of regional communities and the connection among regional communities, and illustrate how regional communities could arouse their activities using the P2P network technology through the simulation result.

1. Introduction

A P2P network technology is advocated in 1998 as the brokerless theory.[1][2][3][4][5][6] So far the P2P networks have been implemented in various aspects on the internet. For example, file exchange services such as Gnutella and Napster, grid computing network (Flet's HIKARI Internet connection service of NTT) [7] and virtual currency control service (Bitcoin by Blockchain Technology.) [8]

We propose to apply the P2P network technology in regional communities because we think that several problems of regional communities are solved by using the P2P network technology.

2. Issues on Regional Communities

2.1. Regional Communities’ Structure

Regional communities have been struggling to make achievements, e.g. a self-directive society, a self-sustaining economic growth and unique business opportunities. But it is difficult for them to lead always their efforts to positive results. We consider that the insufficient results depends on the regional communities’ structures.

So far traditional methods of organizing regional communities are generally a top-down and centralized model, or a bottom-up and flat model. We define the former model as a Client and Server model (C/S model) and the latter one as a P2P pure model.

2.1.1. C/S model

As illustrated in Fig. 1, the C/S model provides communities for a hierarchy form structure. Organization builders, directors or operators take the role of servers and connect themselves to members as clients. The problems of this model are as follows:
- High cost
- Causes community's collapse when the server crashed.
- Causes stop or stagnation of community's activities because of the overflow or the delay of response time on the server when the number of clients increased.

2.1.2. Hybrid model

As illustrated in Fig. 2, the hybrid model is composed by servers and clients as well as the C/S model. The difference between the hybrid model and C/S model is that a server on the hybrid model gives permission its clients to act in substitution for a server. It is used on some file transferring systems, e.g. Napster.

The hybrid model has a problem that the server gets overflowed when the requests from clients increased as well as C/S model.

2.1.3. P2P pure model

As illustrated in Fig. 3, the P2P pure model is flexible because each client works as a connector, and has ability
to build a community characterized by low cost and high scalability.

But this model also has problems that it becomes difficult to manage and maintain clients’ incentives to stay in the community because each client’s incentive is different and various, and causes the community's lawlessness or collapse.

2.2. Non Platform

So far a problem-solving method is a typical way to resolve regional challenges and create community businesses. We found that there are disadvantages to employ the problem-solving method for regional activation because we already recognized the same difficulty in developing software applications not executed on operating systems.

The problem-solving method project contains questions as follows:
- High cost
- Time-consuming
- Overlapping works
- Less quality improvement
- Difficult to share know-hows and collaborate with other projects.

Before we have operating systems, we needed to create a system from the beginning and that was expensive and time-consuming. A knowledge acquired through making a certain system was not able to adapt to other systems.

This problem is all applicable to regional activation. So we classified intersectional functions of the various and unique regional communities’ activities and found that there are common functions among them such as promotion, funding, licensing, presentation and branding. We decided that we gather those common functions and put together to supply regional communities for a tool of regional activation.

From that point of view, the regional communities need to positively introduce the platform model to connect each other and share intersectional functions and make their activities more effective and less costly.

3. Proposing P2P network Model

We propose that the regional communities introduce a P2P network model which has characteristics such as the P2P semi-pure model, incentive-trust-connector and simultaneous participation in different communities.

3.1. P2P Semi-Pure Model

3.1.1. Structure of P2P Semi-Pure Model

Fig. 4 illustrates the component of P2P semi-pure model as follows:
- C/S MODEL A functions as a C/S model.
- The Clients X1 is controlled on the inside of C/S model A.
- PEER A capsuling C/S MODEL A looks as a peer from the other peers (e.g. PEER W) existing on the outside of PEER A.
- X1 and X2 are identical.
- X2 functions as a peer and as a component of the P2P pure model on the outside of PEER A.
- LINK means that the peer situated at either end of the line (e.g. PEER X2 or PEER Z) knows the presence of each other.

As illustrated in Fig. 4, the P2P semi-pure model is a complexed type of the C/S model, the hybrid model and the P2P pure model. In the P2P semi-pure model, a server acts as a peer by being capsuled in a peer, and can participate in a P2P network.

In this case, the server has two faces. One is a face as a peer which can connect to other peers as well as the P2P pure model. The other is a face as a server connecting and controlling its clients as well as the C/S model.

In the latter case, a problem is that the clients capsuled together with its server in a peer will become inactive if the server crashed.

The P2P semi-pure model proposes a unique idea and solution to make the network structure more sustainable.
It allows the clients in the capsuled peer to come out of it and act as an independent peer. In this instance, the client also has two faces. One is as a client connected and controlled by its server. And the other is as a peer which is able to connect with other peers using the P2P network technology.

And as another advantage is that the capsuled server works as a peer and can connect other virtualized servers to compose a massive server network like a grid computing network. When the different servers have connections and work together, synergy effects can be expected. For example, cross-industrial associations often produce innovative outputs.

The P2P semi-pure model also has a load distribution function since the servers and the clients work as peers. We regard that P2P semi-pure model has advantages in cost and scalability in comparison to C/S model and Hybrid model, and is superior to P2P pure model in respect of sustainability and stability.

3.1.2. Simulation of Stability Performance on P2P semi-pure model and P2P Pure Model

As illustrated in Fig. 5, we developed a computer simulation program to measure and compare difference of performance between P2P semi-pure model and P2P pure model in forming virtual communities.

![Fig. 5](image)

**Fig. 5** illustrates the performance comparison of stability of the virtual communities between P2P semi-pure model and P2P pure model. The horizontal axis indicates the number of trials which each of the thirty peers tries a joining and a leaving action to each of ten virtual communities. The vertical axis indicates the number of peers in the largest community, i.e. a community which contains more peers than the others, after each peer's trial of joining and leaving.

In this experiment, we executed the trials ten thousand times and recorded the value every a hundred trials. A parameter “ρ” at the bottom left means the ratio (ρ=λ/μ), i.e. the participation rate (λ) divided by the leave rate (μ). From Fig. 2, P2P semi-pure model is superior to P2P pure model to maintain a stable community.[9]

3.2. Incentive - Trust - Connector

The P2P semi-pure model shows its capabilities such as low costs, high scalabilities and high sustainabilities. Regional communities introducing this model can obtain such merits as follows:

- Incentives of the community members are guaranteed to be fulfilled because the larger the community becomes, more various people, ideas and know-hows are accumulated in the community.
- Trusts of the community to other communities will rise up because the community’s continued existence produces reliability.
- Connectors - The community’s quality becomes better when the number of connectors is increased because the community composed by many peers acquires diversities and sustainabilities. As the result, such community attracts other communities and new participants.

3.3. Simultaneous Participation in Different Communities

Simultaneous participation in different communities, i.e. an action that one person belongs to multiple communities, makes those communities work or relate together. This is one of the important policies of the P2P semi-pure model. There are a lot of patterns to connect multiple communities by means of a few community members. It can make the community bigger. This is what to say, “A friend of a friend of a friend.”

4 Introduction of the SCB Project

4.1. Proposing P2P Platform to regional communities

To make a regional community more effective we propose that the community has the form of the P2P platform by using the P2P semi-pure model.

In concrete terms, as morphing from a person, an activity and a facility to a peer respectively, they are connected with each other to build the P2P platform to host the community’s activities. It is necessary to consider what types of peers are better combined for better activation of the community.
To materialize the P2P Platform for regional activation in the real world, we launched activities using the theory of Social Community Brand [10] at Sojo University SCB Broadcasting Satellite Studio (SCB Studio) in April of 2015.

Social Community Brand, advocated in 2011 by Takashige Hoshiai, is a new theory for regional activation to make communities using the Brokerless Theory such as P2P semi-pure model, simultaneous participation in different communities.

A demonstration experiment is currently in execution at the studio involving people, activities and facility as P2P peers.

4.2. SCB Studio

SCB Studio is established and located in the downtown of Kumamoto City, Japan. It has a broadcast facility on the first floor and “Activity Room” on the second floor.

There are lots of communities in active on the theme such as ICT, agriculture, business start-up, media, sports and voice acting, and the communities are run by the various human resources such as university students, IT engineers, public officers, medical doctors, designers, tour conductors and entrepreneurs.

4.3. Functions of SCB Studio

There are some co-working spaces, shared rented office spaces and rented conference rooms in the same area of SCB Studio. Compared to their service as only renting the space, SCB studio provides users for not only a space but also opportunities as follows:

- Constructing a platform composed of organic link among people, activities and facilities in brokerless, so to speak, it is “Brokerless Link among Regional Resources.”
- Providing useful fundamental resources for the communities such as promotion, funding, licensing, presentation and branding.

The users of SCB Studio can build a platform by not a top-down and centralized model but the P2P semi-pure model with use of “Incentive - Trust - Connector” method and “Simultaneous Participation in Different Communities” method. On this platform, the communities are connected with each other and dissolve regional problems efficiently. In the future, the platforms are likely linked together if such platforms emerged anywhere over the world.

5. Conclusion

Regional communities (especially local communities) are suffered from a shortage of ideas, human resources, funds and social capitals due to uneven distribution of resources. Though it is a new trial to apply the theory of P2P network technology to regional community's activation, we would be truly happy if we see some sprouts of the community with use of the P2P network technology by whoever willingly makes an action for regional activation over the hedges of community, culture and country.

In closing, we would like to thank NOLTA2016 organizing committee members for their fruitful suggestions and comments.

References

Local Noise Sensitivity in Human Photoplethysmogram

Nina Sviridova† and Kenshi Sakai‡

†Meiji Institute for Advanced Study of Mathematical Sciences, Meiji University
4-21-1 Nakano, Nakano-ku, Tokyo 164-8525, Japan
‡ Environmental and Agricultural Engineering Department, Tokyo University of Agriculture and Technology
3-5-8 Saiwai-cho, Fuchu-shi, Tokyo 183-8509, Japan
Email: nina_svr@meiji.ac.jp, ken@cc.tuat.ac.jp

Abstract—Photoplethysmography is one of the widely used techniques to measure biological signals produced by the cardiovascular system. The photoplethysmogram can provide valuable information about cardiovascular system performance. Methods of nonlinear dynamics are recognized to be useful for analyzing the photoplethysmogram; however, their application is often limited by the presence of noise in experimental data. This study sought to investigate the effects of noise on photoplethysmogram dynamics in terms of local noise sensitivity. Wayland test local translation errors were calculated for a photoplethysmogram data set. The results demonstrated that the photoplethysmogram time-delay-reconstructed attractor has a region that is highly affected by noise induction, while the effect of noise on other parts is minor. This finding provides important information for achieving a better understanding of the interaction between noise and photoplethysmogram dynamics, and it may contribute to applied studies utilizing the photoplethysmogram.

1. Introduction

The photoplethysmogram (PPG) is a biological signal produced by the cardiovascular system that is widely used in medical and sports equipment. The PPG signal can be recorded by detecting the near infrared light reflected by vascular tissue following illumination with a LED. The PPG measurement technique is simple and noninvasive, and the PPG is recognized to contain valuable information about cardiovascular system performance [1, 2]. Various studies have demonstrated the usefulness of the PPG analyzed by methods of nonlinear time series analysis, to a wide-range of applications in physical and mental health monitoring [3-6]. However, results obtained from noise-contaminated biological signals, such as the PPG, are often not straightforward to understand or may even be misleading. For example, the Lyapunov exponent calculated numerically from experimental data is not always reliable and might appear positive for non-chaotic data due to the presence of noise in the original signal [7, 8].

Various aspects of noise interaction with complex dynamics signals have been intensively studied [9]. Noise contamination effects, noise filtering, as well as the role of noise as a stochastic force, which allows the creation of realistic models describing natural phenomena and even the possibility of noise-induced chaos, have been investigated in many studies over the last few decades [10-12].

Developing efficient noise filtration techniques or nonlinear time series methods that can perform well for noisy chaotic data may ameliorate noise issues in experimental data. However, conventional noise filtering techniques cannot be efficiently applied to the PPG’s complex dynamics while preserving its unique chaotic characteristics, and modifying existing analysis methods or developing noise-stable analysis methods requires a deep understanding of noise and PPG chaotic dynamics interactions. However, the interaction between noise and complex PPG dynamics and the effect of noise on the PPG are not yet well studied.

In our recent study on noise-induced chaotic models, it was found that noise produces certain effects on chaotic dynamical systems, and a phenomenon that was defined as “local noise sensitivity” was observed. PPG dynamics is recognized to be consistent with chaotic motion [13]. Here we assumed that experimentally obtained PPG signal has a certain amount of noise contamination. It is expected that the PPG dynamics may also demonstrate the presence of a local response to noise induction, which may provide one of the keys necessary for understanding noise-dynamics interactions in the PPG signal.

Understanding the effect of noise on PPG dynamics could significantly contribute to applied studies using the PPG, especially in the area of human health monitoring, as well as to theoretical studies, as it would provide an example of noise produced changes in chaotic motion and may provide the basis for developing noise filtration techniques for chaotic biological data. Therefore, this study sought to investigate the effects produced by noise on the PPG in terms of its local noise sensitivity.

2. Methods and Materials

2.1. Photoplethysmogram data collection

The PPG data were collected by a finger PPG recorder from healthy 19- to 27-year old volunteers among Tokyo University of Agriculture and Technology (TUAT) students. Experimental data collection was approved by TUAT.
authorities. Written informed consent was given to participants prior to the experiment. At the time of the study all subjects were healthy non-smokers, physically active to similar levels, were not taking any medication, and all of them declare no history of heart disease.

For each participant the measured period was 5 min with 5 msec sampling steps. For all data collection sessions, a BACS (Computer Convenience, Inc.) PPG sensor was located on the right forefinger. According to the recommended settings for data collection and significant factors that may affect PPG measurements described previously [14], all measurements were done with the subject in a relaxed sitting position in a room with temperature, noise and vibration control. Prior to the test, each test subject was asked to rest for 5 min under quiet conditions in the laboratory room with the test site uncovered.

2.2. Local Noise Sensitivity

In our previous works [6, 13] the Wayland test translation error [15] was successfully utilized for detecting determinism in the PPG signal obtained in a reference environment and under noise exposure. The translation error appeared to be quite a useful index for characterizing the chaotic properties of the PPG signal, and it provided us with quantitative information regarding the determinism of the investigated PPG signal and the smoothness of its reconstructed attractor. Previously only the translation error averaged over the reconstructed trajectory was applied to the PPG. In this study we calculated the Wayland test local translation error (LTE) in every point of the data set using the following formula [15]:

\[ e_{loc} = \frac{1}{k + 1} \sum_{j=0}^{k} \frac{\|v_j - \langle v \rangle\|^2}{\|\langle v \rangle\|^2}, \]

where \( v_{ij} \) are translation vectors, \( \langle v \rangle \) is the average of \( v_j, x_j \) (\( j=1,2,\ldots,k \)) are \( k \) nearest neighbors to the fixed and arbitrary chosen point \( x_0 \) on the reconstructed trajectory and \( v_j \) are projections of \( x_j \).

Previously, a local noise sensitivity phenomenon was detected in chaotic dynamical systems by utilizing LTE on examples of the Lorenz and the Rössler systems in the chaotic regime. As the concept of local noise sensitivity is new, it will be briefly described below in an example of the chaotic Lorenz model.

3. Results

Experimentally obtained biological data, such as the PPG, are inevitably noise contaminated, which may complicate drawing conclusions from the analysis results. Therefore, to analyze the effects of noise on the PPG dynamics, the well-known Lorenz model in the chaotic regime and its local noise sensitivity were utilized for comparative investigation.

3.1. The chaotic Lorenz model

The Lorenz model, which is one of the typical examples of models generating chaos, is described by the following system of equations [16]:

\[
\begin{align*}
\dot{x} &= -\sigma(x - y), \\
\dot{y} &= \rho x - y - xz, \\
\dot{z} &= xy - \beta z,
\end{align*}
\]

where system parameters \( \sigma=10, \rho=28, \beta=8/3 \) correspond to the chaotic regime [16]. This system was solved numerically with the 4th order Runge-Kutta method.

The LTE was calculated in each data point for the data obtained as numerical solutions of the chaotic Lorenz system above, and for noise-induced Lorenz data. Results were plotted along a reconstructed attractor as shown in Fig. 1. Noise, whose components are uniformly distributed random numbers, with 7% of its amplitude, was induced on the Lorenz data. For clarity in the graph the threshold for LTE was chosen as 0.5 since this or any higher LTE value would correspond to non-deterministic flow. Fig. 1 (a) corresponds to the noise-free case, i.e., to the data calculated from the above described model. Fig. 1 (b) shows the case of 7% noise induced on the Lorenz data.

**FIG. 1** The distribution of local translation error along the time-delay-reconstructed chaotic Lorenz attractor: (a) noise-free and (b) 7% noise induction case.

As seen from Fig. 1 (a) and (b) with noise induction, the number of high values for the LTE increased; moreover,
high values were not distributed evenly along the trajectory but were concentrated in certain regions of the reconstructed attractor. This phenomenon can be observed in chaotic models and it is called local noise sensitivity.

3.2. Photoplethysmogram

The LTE was calculated for the PPG at each data point. LTE distribution along the PPG trajectory is shown in Fig. 2 (a). As seen in Fig. 2, only several areas have a clearly observable concentration of high LTEs, while other parts of the trajectory have LTEs close to zero. However, although we assumed noise contamination of the PPG data, the actual amount of noise as well as the types of noise induced during data collection are unknown. This makes it difficult to draw unambiguous conclusions from results shown in Fig. 2 (a). As shown in Fig. 1 for the Lorenz model, with subsequent noise increase, the high LTE values area expanded. To test whether the PPG’s LTE distribution would significantly change in response to further noise increases, additive noise, as in the Lorenz case, was induced on the PPG time series. The LTE distribution for noise induced PPG is shown in Fig. 2 (b) and (c) for noise with the original amplitude and with a two-fold increase in the amplitude value, respectively.

The global Wayland test translation errors were 0.0561, 0.0936, and 0.1841 for the original, noise-induced and double amplitude noise-induced PPG data.

As seen in Fig. 2, similar to the chaotic Lorenz case, high LTE values tended to concentrate in certain regions of the PPG reconstructed attractor; however, the character of the distribution, the number of high LTE concentration areas, as well as distribution changes under further noise induction, differed from the Lorenz case.

It is important to notice that the noise amplitude values were chosen in consideration with the Lorenz and PPG data scales, as well as the PPG time series finiteness and resolution.

4. Discussion

The main objective of this study was to investigate the effects produced by noise on the PPG dynamics in terms of local noise sensitivity.

Even though the actual noise level in the PPG is unknown, it was assumed that experimentally obtained PPG data must contain a certain amount of noise. Under this assumption and taking into account that the PPG dynamics is consistent with the definition of chaotic motion, the presence of local noise sensitivity, which was previously demonstrated for the chaotic Lorenz model, was investigated.

Similar to the Lorenz case, the distribution of the LTE along the PPG trajectory demonstrated the presence of local regions on the PPG attractor with high value LTEs concentrated areas (Fig. 2 (a)). However, local noise sensitivity is not an exclusive reason for high LTE values, and therefore it was important to observe the changes in the distribution in response to further noise induction. As demonstrated for the chaotic Lorenz system (Fig. 1), in noise sensitive regions, the area with high LTE values appeared to be spreading as the noise increased. For the original PPG time series four regions on the attractor demonstrated concentrations of high value LTEs: three rapidly bending parts and the spiral part of the trajectory. However, subsequent noise induction on the PPG time series (Fig. 2 (b-c)) only resulted in increased LTE values over the spiral region of the attractor, while the distribution at three other regions remained essentially unchanged. The concentration of high value LTEs in the spiral region of the attractor and its distribution dynamics in response to increased induced noise indicates the presence of a local noise sensitivity phenomenon in the spiral region of the PPG time-delay-reconstructed attractor. However, conclusions regarding the presence of local noise...
sensitivity for the other three regions with high value LTEs cannot be drawn unambiguously and further investigation of the response of these regions to noise induction is required. It is important to notice that in this study only the simplest case of additive noise with uniform random distribution was analyzed, and therefore, future study of the effects of various types of noise may provide the key for understanding the noise-dynamics interaction of the other three regions.

Local reaction of the spiral part of the trajectory expressed by the increase of LTE values can be partially explained by taking into consideration the fact that in spiral and bending areas of an attractor reconstructed from a discrete time series, data point density is higher compared with other parts of the trajectory. Therefore, the presence of noise, whose scale is comparable with the original data, can create a significant disturbance in the trajectory. In this case due to the high data points concentration, relatively remote noise-contaminated data points may replace closer ones and act as nearest neighbors. Due to these false neighbors, not only the LTE may have high values even in apparently deterministic data, but it may also affect various nonlinear time series analysis methods that rely on nearest neighbor searches or the smoothness of the reconstructed trajectory.

Additionally, apparent differences in the type of LTE distribution and the number of regions with high LTE between the chaotic Lorenz case and the PPG may be partially explained by topological differences between the Lorenz layering type attractor and the PPG attractor, which can be recognized as a folded band.

The presence of a noise sensitive region on the time-delay reconstructed PPG attractor discovered in this study can improve our understanding of the effect of noise on PPG dynamics, and it is expected to contribute to further applied studies of the PPG signal.

5. Conclusions

In this study the local Wayland test translation error was utilized for investigation of the effect of noise on the PPG chaotic dynamics. Results demonstrated that the PPG display local noise sensitivity, which indicates that noise present in the experimental PPG data does not affect it evenly, and while noise may cause considerable effects, they only occur in specific regions. The noise-sensitive portion of the PPG attractor, where the LTE reached high values and its area increased with subsequent noise induction, was represented by the spiral part of the reconstructed trajectory.

Discovery of local noise sensitivity in the human PPG can significantly contribute to various applications in the field of human health monitoring. Additionally, it is expected that further comparative investigation of the PPG noise sensitivity may lead to the development of methods for estimating noise levels in experimental PPG data.

Acknowledgments

This work was partially supported by JSPS Grant-in-Aid No. 25660204 and 16K15016.

References

Security Analysis of a Chaos Based Random Number Generator

Salih Ergün

†TÜBİTAK-Informatics and Information Security Research Center
PO Box 74, 41470, Gebze, Kocaeli, Turkey
Email: salih.ergun@tubitak.gov.tr

Abstract — This paper introduces security analysis of a chaos based random number generator (RNG). An attack system is proposed to discover the security weaknesses of the chaos-based RNG. Convergence of the attack system is proved using auto-synchronization scheme (synchronization with unknown parameters). Secret parameters of the RNG are recovered from a scalar time series where the only information available are the structure of the RNG and a scalar time series observed from the chaotic oscillator. Simulation and numerical results verifying the feasibility of the attack system are given. It is shown that deterministic chaos itself cannot be pointed out as the source of randomness.

1. Introduction

Over the last decades there has been an increasing emphasis on using tools of information secrecy. Certainly, random number generators (RNGs) have more prominently positioned into the focal point of research as the core component of the secure systems [1]. Although many people are even unaware that they are using them, we use RNGs in our daily business. If we ever drew money from a bank, ordered goods over the internet with a credit card, or watched pay TV we have used RNGs. Public/private key-pairs for asymmetric algorithms, keys for symmetric and hybrid cryptosystems, one-time pad, nonces and padding bytes are created by using RNGs [4].

Being aware of any knowledge on the design of the RNG should not provide a useful prediction about the output bit sequence. Even so, fulfilling the requirements for secrecy of cryptographic applications using the RNG dictates three secrecy criteria as a “must”: 1. The output bit sequence of the RNG must pass all the statistical tests of randomness; 2. The previous and the next random bit must be unpredictable [2] and; 3. The same output bit sequence of the RNG must not be able to be reproduced [3].

An important principle of modern cryptography is the Kerckhoff’s assumption [1], states that the overall security of any cryptographic system entirely depends on the security of the key, and assumes that all the other parameters of the system are publicly known. Security analysis is the complementary of cryptography. Interaction between these two branches of cryptography forms modern cryptography which has become strong only because of security analysis revealing weaknesses in existing cryptographic systems.

Although the use of discrete-time chaotic maps in the realization of RNG has been widely accepted for a long period of time, it has been shown during the last decade that continuous-time chaotic oscillators can also be used to realize RNGs [5, 6]. In particular, a so-called RNG based on a continuous-time chaotic oscillator has been proposed in [5]. In this paper we target the RNG reported in [5] and further propose an attack system to discover the security weaknesses of the targeted system. The strength of a cryptographic system almost depends on the strength of the key used or in other words on the difficulty for an attacker to predict the key. On the contrary to recent RNG design [6], where the effect of noise generated by circuit components is analyzed to address security issue, the target random number generation system [5] pointed out the deterministic chaos itself as the source of randomness.

The organization of the paper is as follows. In Section 2 the target RNG system is described in detail; In Section 3 an attack system is proposed to cryptanalyze the target system and its convergence is proved; Section 4 illustrates the numerical results with simulations which is followed by concluding remarks.

2. Target System

Chaotic systems are categorized into two groups: discrete-time or continuous-time, respectively regarding on the evolution of the dynamical systems. The double-scroll chaotic system is considered as one of the most famous continuous-time chaotic system that have ever been introduced, many designs of which were proposed starting from the use of a structure similar to Chua’s oscillator. Double-scroll-like attractor which is used as the core in target random number generation system [5] is obtained from a simple model which is expressed by the Eqn. 1.
Given double-scroll chaotic system is single-parameter-controlled where \( a \) is the only parameter which contributes to the chaotic dynamics. The equations in 1 generate chaos for the single-parameter \( a \) over a wide range \((0.48 < a < 1)\) which points out that there is enough clearance for the latter. For analyzing the target RNG, the chaotic attractor 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.

\[
\begin{align*}
  \dot{x}_1 &= y_1 \\
  \dot{y}_1 &= z_1 \\
  \dot{z}_1 &= -a_1 x_1 - a_1 y_1 - a_1 z_1 + a_1 \text{sgn}(x_1)
\end{align*}
\]  

(1)

Target RNG is based on periodic sampling of chaotic oscillator. Periodic samples of the state variable \( x_1 \) in Equation 1 were used. These samples are obtained at the rising edges of an external periodical pulse signal, that is at times \( t \) satisfying \( w t \mod 2\pi = 0 \) where \( w \) is the frequency of the pulse signal.

In target RNG, the distribution of periodically sampled \( x_1 \) values was initially examined to determine appropriate sections where the distribution looks like random signal. For different values of \( a_1 \) given in Equation 1, various sections were determined where the distribution of \( x_1 \) has two regions. Following this direction, bit sequence \( S_{(top)} \) and \( S_{(bottom)} \) were generated for \( a = 0.666 \) from regional \( x_1 \) values for regional thresholds according to the Equation 2:

\[
\begin{align*}
  S_{(top)} &= \text{sgn}(x_1i - q_{top}) & \text{when } x_1i \geq q_{middle} \\
  S_{(bottom)} &= \text{sgn}(x_1i - q_{bottom}) & \text{when } x_1i < q_{middle} \\
  S_{(xor)} &= S_{(top)} \bigotimes S_{(bottom)}
\end{align*}
\]  

(2)

where \( x_1i \)'s are the values of \( x_1 \) at the 1-dimensional section, \( q_{top} \) and \( q_{bottom} \) are the thresholds for top and bottom distributions, respectively, \( q_{middle} \) is the boundary between the distributions and \( \bigotimes \) is the exclusive-or operation used to generate random bit streams. It should be noted that, anyone who knows the chaotic signal output can reproduce the same output bit sequence \( S_{(xor)} \).

Numerical and experimental results verifying the correct operation of the proposed RNG were presented in [5] such that numerically generated binary sequences fulfill FIPS-140-2 test suite [7] while TRNG circuit fulfill the NIST-800-22 statistical test suite [8]. It should be noted that, the target random number generation system [5] satisfies the first secrecy criteria, which states that “TRNG must pass all the statistical tests of randomness.”

![Figure 1: Largest CLEs as a function of coupling strength c.](image)

3. Attack System

After the seminal work on chaotic systems by Pecora and Carroll [9], synchronization of chaotic systems has been an increasingly active area of research [10]. In this paper, convergence of attack and target systems is numerically demonstrated using auto-synchronization scheme [11] which is known as synchronization of chaotic systems with unknown parameters. In order to provide security analysis of the target random number generation system an attack system is proposed which is given by the following Eqn. 3:

\[
\begin{align*}
  \dot{x}_2 &= y_2 \\
  \dot{y}_2 &= z_2 + c(y_1 - y_2) \\
  \dot{z}_2 &= -a_2 x_2 - a_2 y_2 - a_2 z_2 + a_2 \text{sgn}(x_2) \\
  \dot{a}_2 &= -y_1(y_1 - y_2)
\end{align*}
\]  

(3)

where \( c \) is the coupling strength between the target and attack systems \( a_2 \) is the unknown control parameter of the target system to be estimated. The only information available are the structure of the target random number generation system and a scalar time series observed from \( y_1 \).
In this paper, we able to construct the attack system expressed by the Eqn. 3 that synchronizes \( x_2 \rightarrow x_1 \) for \( t \rightarrow \infty \) where \( t \) is the normalized time.

We define the error signals as \( e_x = x_1 - x_2 \), \( e_y = y_1 - y_2 \) and \( e_z = z_1 - z_2 \) where the aim of the attack is to design the coupling strength such that \( |e(t)| \rightarrow 0 \) as \( t \rightarrow \infty \).

The auto-synchronization of attack and target systems is verified by the conditional Lyapunov Exponents (CLEs), and as firstly reported in [9], is achievable if the largest CLE is negative. In Fig.1, largest CLE graph is drawn as a function of coupling strength \( c \) while the normalized time. As drawn in the figure, when \( c \) is greater than 0.4 then the largest CLE is negative and hence identical synchronization of target and attack systems starting with different initial conditions is achieved and stable [9]. However for \( c \) is equal to or less than 0.4, largest CLE is positive and identical synchronization is unstable.

As shown in Fig.2, the attack system converges to the parameter \( a_1 \) of the target system and auto-synchronization is achieved in less than 70\( t \). Log \( |e_x(t)| \), Log \( |e_y(t)| \) and Log \( |e_z(t)| \) are shown in Fig.3 for \( c = 2 \), where the synchronization effect is better than that of \( c = 0.4 \).

4. Numerical Results

We numerically demonstrate the proposed attack system using a 4\textsuperscript{th}-order Runge-Kutta algorithm with fixed step size and its convergence is illustrated in Fig.3. Numerical results of \( x_1 - x_2 \), \( y_1 - y_2 \) and \( z_1 - z_2 \) are also given in Fig. 4, Fig. 5, and Fig. 6, respectively illustrating the unsynchronized behavior and the synchronization of target and attack systems.

It is observed from the given figures that, auto-synchronization is achieved and stable. As shown by black lines in these figures, no synchronous phenomenon is observed before 70\( t \). In time, the proposed attack system converges to the target system and identical synchronization is achieved where colored lines depict synchronized behaviors of chaotic states in Fig. 4, Fig. 5, and Fig. 6, respectively.

5. Conclusions

In this paper, we propose a numerical attack on a chaos based random number generator (RNG). An attack system is introduced to discover the security weaknesses of the chaos-based RNG and its convergence is proved using auto-synchronization scheme. Although the only information available are the structure of the target RNG and a scalar time series ob-
Figure 5: Numerical result of $y_1 - y_2$ illustrating the unsynchronized behavior and the synchronization of target and attack systems.

Figure 6: Numerical result of $z_1 - z_2$ illustrating the unsynchronized behavior and the synchronization of target and attack systems.

served from the target chaotic system, identical synchronization of target and attack systems is achieved and hence output bit streams are synchronized. Moreover, it is shown that secret parameters can be recovered by using auto-synchronization scheme. Simulation and numerical results presented in this work not only verify the feasibility of the proposed attack but also encourage its use for the security analysis of the other chaos based RNG designs.

References


Application of a Bivariate Fractal Interpolation Surface to an Analysis of Perspective Painting Images

Kazuya Ozawa‡, Hiroyuki Yamada‡, Hideo Nakano‡ and Hideaki Okazaki‡

‡Graduate School of Engineering, Shonan Institute of Technology
1–1–25, Tsujidounishikaigan, Fujisawa-shi, Kanagawa Prefecture 251-8511, Japan
‡Faculty of Engineering, Shonan Institute of Technology
1–1–25, Tsujidounishikaigan, Fujisawa-shi, Kanagawa Prefecture 251-8511, Japan
Email: 13a6029@sit.shonan-it.ac.jp, 16t2009@sit.shonan-it.ac.jp, nakano@sc.shonan-it.ac.jp, okazaki@sc.shonan-it.ac.jp

Abstract—In this paper, a lemma and two theorems for a bivariate fractal interpolation surface (BFIS) generated by an iterated function system (IFS) with an individual vertical scaling factor are introduced. Using an affine transformation and the BFIS, a foundation of analysis of perspective painted images, including images from an arbitrary vantage point together with different sfumato effects, is provided. Finally, this analysis of perspective for painted images is applied to the Virgin section of Annunciation by Leonardo da Vinci, and followed by concluding remarks.

1. Introduction

In [1], a lemma and two theorems for a bivariate fractal interpolation surface (BFIS) generated by an iterated function system (IFS) with an individual vertical scaling factor, are provided. For broader application, use of the individual vertical scaling factor for the BFIS is emphasized.

Computer image analysis, such as image warping methods [2], [3] can help art historians to answer much-debated questions; for example, is the geometry of Masaccio’s Trinity correct, or what is the shape of the dome in Raphael’s School of Athens? Other interesting questions for the spatial structure of paintings (e.g. linear perspective, anamorphosis, aerial or atmospheric perspective, and sfumato used “to tone down” in a painting or drawing [4]) can also be addressed. Since interpolation methods are indispensable for image warping, the BFIS with an individual vertical scaling factor is expected to be effective for such computer image analysis. In this paper, we consider applying the BFIS to analysis of perspective paintings, such as linear perspective, and sfumato that is often used in connection with the works of Leonardo da Vinci. For purpose of analysis, we focus upon Leonardo’s Annunciation [7], [8], [9].

First, we introduce a lemma and two theorems for an IFS generating a BFIS with the individual vertical scaling factor shown in [1]. Secondly, we provide a foundation of analysis for perspective painting images, including images from an arbitrary point of view together with different sfumato effects, using an affine transformation and the BFIS. We then apply this analysis of perspective painted images to a key section of Annunciation, specifically the area containing the Virgin. Finally, we summarize the concluding remarks.

In the following, let Z be a set of integer numbers, \( \mathbb{Z} \), be a set of non-negative integer numbers, and \( \mathbb{R} \) be a set of real numbers.

2. Fractal interpolation surface on the rectangular grid points

This section introduces the lemma and two theorems for a bivariate fractal interpolation surface (BFIS) generated by an iterated function system (IFS) with the individual vertical scaling factor shown in [1].

Let \( I = [0, 1] \subset \mathbb{R} \) and \( D = [0, 1]^2 \subset \mathbb{R}^2 \). And let \( N > 1, M > 1, N, M \in \mathbb{Z}^+ \) be two positive integers. Set \( x_i = \frac{i}{N}, y_j = \frac{j}{M}, i = 0, \ldots, N, j = 0, \ldots, M \). We obtain a set of rectangular grid points \( \{(x_i, y_j) \in \mathbb{R}^2 | i = 0, \ldots, N, j = 0, \ldots, M \} \) of \( D \). Let \( \psi(x, y) : D \subset \mathbb{R}^2 \rightarrow \mathbb{R} \) be a continuous function defined on \( D \), and \( z_{ij} = \psi(x_i, y_j) \). Then

\[
\{(x_i, y_j, z_{ij}) \in \mathbb{R}^3 | i = 0, \ldots, N, j = 0, \ldots, M \}
\]

is a set of data points (interpolating points) in \( \mathbb{R}^3 \).

Let \( u_i(x) \times v_j(y) : D \rightarrow [x_{i-1}, x_i] \times [y_{j-1}, y_j] \subset D, i = 1, \ldots, N, j = 1, \ldots, M, \) as follows:

\[
\left( \begin{array}{c}
  u_i(x) \\
  v_j(y)
\end{array} \right) = \left( \begin{array}{c}
  (\frac{1}{M})^{\frac{(i-1)M}{2}} + \frac{i\sigma(k)}{N} \\
  (\frac{1}{M})^{\frac{(j-1)M}{2}} + \frac{j\sigma(k)}{N}
\end{array} \right)
\]

where \( \sigma(k) = k \mod 2 \).

Let \( f_{ij}(x, y) : [x_{i-1}, x_i] \times [y_{j-1}, y_j] \subset D \rightarrow \mathbb{R}, i = 1, \ldots, N, j = 1, \ldots, M, \) be the piecewise polynomial interpolation function, as follows:

\[
f_{ij}(x, y) = \sum_{(k, l) \in I \times I \setminus (i-1, j-1)} z_{kl} \cdot \Phi_{\Delta_{ij}}(u_i(x) - 1, v_j(y)^{-1})
\]

where \( \Delta_{ij} = (N\sigma(k), M\sigma(l)) \). The bivariate functions \( \Phi_{N,M}, \Phi_{0,0}, \Phi_{0,M}, \Phi_{N,0} : D \rightarrow I \) are defined as

\[
\Phi_{N,M}(x, y) = xy, \Phi_{0,0}(x, y) = (1-x)(1-y), \Phi_{0,M}(x, y) = (1-x)y, \Phi_{N,0}(x, y) = x(1-y).
\]
Define mappings \( w_{ij} : D \times \mathbb{R} \to \mathbb{R}^3, \) \( i = 1, \ldots, N, j = 1, \ldots, M, \) as follows:

\[
w_{ij} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} u_i(x) \\ v_j(y) \\ h_{ij}(x, y, z, d_{ij}) \end{pmatrix},
\]

\[
= \begin{pmatrix} u_i(x) \\ v_j(y) \\ f_{ij}(u_i(x), v_j(y)) + g_{ij}(u_i(x), v_j(y), z, d_{ij}) \end{pmatrix}, \tag{4}
\]

where \( d_{ij} \) is a free parameter, which is considered the vertical scaling factor. Here, \( g_{ij}(x, y, z, d_{ij}) : [x_{j-1}, x_j] \times [y_{j-1}, y_j] \subset D \times \mathbb{R} \times [0, 1] \to \mathbb{R}, \) \( i = 1, \ldots, N, j = 1, \ldots, M, \) is described, as follows:

\[
g_{ij}(x, y, z, d_{ij}) = d_{ij}z - \sum_{(k,l)\in[-1,1] \times [-1,1]} d_{k,l}z_{k,l} \cdot \Phi_{k,l}(u_i(x)^{-1}, v_j(y)^{-1}), \tag{5}
\]

**Lemma 1** \( h_{ij}(x, y, z, d_{ij}) : D \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}, \) \( i = 1, \ldots, N, j = 1, \ldots, M, \) has the following relations such that

\[
|h_{ij}(x_1, y_1, z_1, d_{ij}) - h_{ij}(x_2, y_2, z_2, d_{ij})| \leq c \cdot |x_1 - x_2| + c \cdot |y_1 - y_2|, \tag{6}
\]

\[
|h_{ij}(x_1, y_1, z_1, d_{ij}) - h_{ij}(x_1, y_2, z_2, d_{ij})| \leq c \cdot |z_1 - z_2|, \tag{7}
\]

\[
c_{ij} = 4 \cdot \max_{(k,l)\in[-1,1] \times [-1,1]} \max_{1 \leq s \leq 2} \max_{1 \leq t \leq 2} |z_{s,t} - d_{ij}z_{s,t}|,
\]

\[
c = \max_{1 \leq i \leq N, 1 \leq j \leq M} c_{ij},
\]

\[
|d_{ij} \cdot |z_1 - z_2|/|d_{ij}| |z_1 - z_2|, f or x \in I, y \in I, z_1, z_2 \in \mathbb{R}, \tag{8}
\]

\[
0 \leq d_{ij} < 1.
\]

**Theorem 1 (IFS Theorem)** Let \( N > 1, M > 1, N, M \in \mathbb{Z}_+ \) be two positive integers. Let \( D \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \) denote the mappings (or IFS) of Eq.(4), associated with the data set of Eq.(1). Let the vertical scaling factor \( d_{ij} \) obey \( 0 \leq |d_{ij}| < 1 \) for \( i = 1, \ldots, N, j = 1, \ldots, M. \) Then there is a metric \( d_m \) on \( \mathbb{R}^3, \) equivalent to the Euclidean metric, such that IFS of Eq.(4) is hyperbolic with respect to \( d_m. \) In particular, there is a unique non-empty compact set, or an attractor: \( G \subset \mathbb{R}^3 \) such that

\[
G = \bigcup_{i,j}^{N,M} w_{ij}(G) \tag{9}
\]

**Theorem 2 (FIS Theorem)** Let \( N > 1, M > 1, N, M \in \mathbb{Z}_+ \) be two positive integers. Let \( D \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R} \) denote the mappings (or IFS) of Eq.(4), associated with the data set of Eq.(1). Let the vertical scaling factor \( d_{ij} \) obey \( 0 \leq |d_{ij}| < 1 \) for \( i = 1, \ldots, N, j = 1, \ldots, M, \) so that the IFS of Eq.(4) is hyperbolic. Let \( G \) denote the attractor of the IFS. \( G \) is the graph of a continuous function of \( \phi(x, y) : D \subset \mathbb{R}^2 \to \mathbb{R}, \) which interpolates the data of Eq.(1). That is,

\[
\hat{G} = \{(x, y, \phi(x, y)) : (x, y) \in D, \}
\]

\[
where \( \phi(x, y) = \phi(x_0, y_0), \) for \( i = 1, \ldots, N, j = 1, \ldots, M. \) (11)

1. If any \( i, j, d_{ij} \) are identical to each other; i.e. \( d = d_{ij}, \) then \( G = \hat{G}. \)

2. The attractor \( G \) of Eq.(4) includes the set of data points of Eq.(1).

3. The attractor \( G \) of Eq.(4) depends continuously on \( d_{ij}. \)

### 3.2. Affine transformation [6], pp. 49 - 53

To obtain an image of the painting from an arbitrary vantage point, we define an affine map \( A \) as follows:

\[
A \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} = \begin{pmatrix} r_1 \cos \theta_1 & r_2 \cos \theta_2 \\ r_1 \sin \theta_1 & r_2 \sin \theta_2 \end{pmatrix} \begin{pmatrix} x^0 \\ y^0 \end{pmatrix} \tag{12}
\]

Here let \( N_0 > 1, M_0 > 1, N_0, M_0 \in \mathbb{Z}_+ \) be two positive integers. We obtain another set of rectangular grid points \( \{(x_i^0, y_j^0) \in \mathbb{R}^2 \mid 0, \ldots, N_0, j = 0, \ldots, M_0\} \) of \( D. \)

Let \( \phi(x^0, y^0) : D \subset \mathbb{R}^2 \to \mathbb{R} \) be a continuous function defined on \( D, \) and \( \phi(x_i^0, y_j^0) \). Then

\[
\{(x_i^0, y_j^0, \phi(x_i^0, y_j^0)) \in \mathbb{R}^3 \mid 0, \ldots, N_0, j = 0, \ldots, M_0\} \tag{13}
\]

is a set of data points (interpolating points) in \( \mathbb{R}^3. \)
If the source image $\phi$ of Eq.(13) defined over a $(x^0, y^0)$ coordinate system undergoes an affine transformation (i.e. linear perspective) from an arbitrary vantage point to produce a destination image $\phi'$ of Eq.(1) defined over an $(x, y)$ coordinate system, this affine transformation (of the coordinates) is expressed as Eq.(12):

$$[x, y]^T = A [x^0, y^0]^T.$$  

(14)

3.3. The images from another vantage point together with different sfumato effects

Subsequently, we focused on the right area of Annunciation, specifically the Virgin section of the Annunciation image. The source image $\phi$ : the Virgin section of the Annunciation image, and $\phi'$ : the edge image of $\phi$ are illustrated in Fig. 1, and Fig. 3, respectively. Using the affine transformation of Eq.(14) with the parameter : $r_1 = 1, r_2 = 47/50$, $\theta_1 = \pi/180$, and $\theta_2 = 0$, we identify the destination images from about a forty-five degree angle to the right, $\psi$, 256$^3$ RGB levels of Fig. 2, and $\psi'$ from the source images $\phi$, 256$^3$ RGB levels of Fig. 1 and $\phi'$, 256 gray levels of Fig. 3. Further using the mappings of Eq.(4), and an algorithm similar to Algorithm 8.2 [6], pp.84 - 91, on MATLAB, we illustrate the examples of sfumato effects for the painted images, 256$^3$ RGB levels, size $(N, M)$, from the modified image $\psi$, size $(N', M')$: $(N, M) \rightarrow (N', M')$. From the FIS generated by Eq.(4), the number of iterations: $3 \times 10^5$, and the composite images of the Virgin consisting of Red, Green, and Blue color reconstructed images, are illustrated in Fig 4: $(d_{ij} = 0.3$, for the edge of $\psi')$, and in Fig. 5: $(d_{ij} = 0.09$, for the edge of $\psi')$. In each color image, $d_{ij} = 0.03$ for the other region except for the edge of $\psi'$, respectively. Comparing the viewpoints and the sfumato effects between Fig. 1, Fig. 4, and Fig. 5, we can see the difference in the Virgin’s facial expressions. For purposes of comparison, an existing technology with similar properties to those of this study include "atmospheric perspective effect enhancement" [5]. A study of the same painting would be the ideal; however, the authors were unable to discover an analysis of Annunciation using this technology. On the other hand, a research study, Atmospheric Perspective Effect Enhancement of Landscape Photographs Through Depth-Aware Contrast Manipulation[5], was accessed and provided data which could be used for a comparison (see Table 1).

4. Concluding Remarks

1. We have introduced the lemma and theorems to establish the BFIS with an individual vertical scaling factor for every set of rectangular grid data, and have provided the foundation for analysis of perspective painted images, including images from arbitrary vantage points together with different sfumato effects, using an affine transformation and the BFIS.

2. We have shown examples of the sfumato effects, in particular, in Fig. 4 and in Fig. 5. The effect of the individual vertical scaling factor (i.e. the strength of the sfumato) was demonstrated such that the eyes and the mouth in Fig. 5 are more chiseled than those in Fig. 4, when magnified.

References


[9] Virtual Uffizi Gallery “The mystery of Leonardo’s Annunciation,” ‘‘: mistakes or virtuosities?,”

https://www.virtualuffizi.com/ the-mystery-of-leonardo %E2%80%99s-annunciation %3A-mistakes-or-virtuosities%3F.html
Table 1: Comparison of our work and an existing technology [5]

<table>
<thead>
<tr>
<th>contrast</th>
<th>this study</th>
<th>depth-aware contrast manipulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>individual vertical</td>
<td></td>
<td>inter-contrast and</td>
</tr>
<tr>
<td>scaling factor control</td>
<td></td>
<td>intra-contrast control</td>
</tr>
</tbody>
</table>

Figure 1: The Virgin section of *Annunciation* image $\phi$ : size $(N_0, M_0) = (2031, 1856)$

Figure 2: The Virgin section of *Annunciation* image $\psi$ : size $(N_0, M_0) = (2031, 1892)$

Figure 3: Edge image $\phi'$ of Fig. 1 : size $(N_0, M_0) = (2031, 1892)$. The lightest sfumato is used to effect the graininess of the figure (NB apart from the face).

Figure 4: A reconstruction of the Virgin image from the FIS (1) : size $(N', M') = (2031, 1892)$. The lightest sfumato is used to effect the graininess of the figure (NB apart from the face).

Figure 5: A reconstruction of the Virgin image from the FIS (2) : size $(N', M') = (2031, 1892)$. The second lightest sfumato is used to effect the graininess of the figure (NB apart from the face).
Automatic Melody Generation considering Chord Progression using Genetic Algorithm

Yudai WAKUI, Yoshinori HATORI and Yuko OSANA

School of Computer Science, Tokyo University of Technology
1404-1 Katakura, Hachioji, Tokyo, 192-0982, Japan Email: osana@stf.teu.ac.jp

Abstract—In this paper, we propose an automatic melody generation system considering chord progression. In the proposed system, chord progression and rhythm sequence are generated randomly, and the pitch is assigned to each note using genetic algorithm. We carried out a series of computer experiments, and we confirmed that melodies can be generated by the proposed system.

1. Introduction

Since the first approach to the automatic composition in 1957, a lot of methods for automatic composition have been proposed[1]–[6]. As one of these methods, we have proposed the automatic melody generation method using N-gram model and genetic algorithm[5]. In this method, the features on sample melodies are trained using N-gram models[8] per melody blocks such as verse, bridge and chorus. And melodies that have similar features to trained sample melodies can be generated using genetic algorithm[7]. However, in most of these systems, melodies are generated based on a sample melody or sample melodies given by users.

In this paper, we propose an automatic melody generation system considering chord progression. In the proposed system, chord progression and rhythm sequence are generated randomly, and the pitch is assigned to each note using genetic algorithm.

2. Automatic Melody Generation System considering Chord Progression by Genetic Algorithm

Here, the proposed automatic melody generation system considering chord progression by genetic algorithm is explained.

The proposed system has four phases: (1) condition input, (2) chord progression generation, (3) rhythm sequence generation and (4) pitch assignment by genetic algorithm.

2.1. Condition Input

First, a user input following conditions.

- The length of note used mainly is chosen from quarter note and eighth note, or eighth note or sixteenth note.
- Key is selected.
- The length of each section is selected from 4, 8 or 16 bars.
- Whether auftakt is used or not is selected.
- The distribution of difference between two sounds are determined.

2.2. Chord Progression Generation

2.2.1. Motif Transition

In the proposed system, each motif consist of two bars. First, motif transition are determined.

If a motif has four bars, motif transition is selected from

- motif A \rightarrow motif A
- motif A \rightarrow motif B

randomly.

If a motif has eight bars, motif transition is selected from

- motif A \rightarrow motif A \rightarrow motif A \rightarrow motif A
- motif A \rightarrow motif A \rightarrow motif B
- motif A \rightarrow motif B \rightarrow motif A \rightarrow motif B
- motif A \rightarrow motif A \rightarrow motif B \rightarrow motif B
- motif A \rightarrow motif A \rightarrow motif B \rightarrow motif C
- motif A \rightarrow motif B \rightarrow motif A \rightarrow motif C

randomly. If a motif has 16 bars, motif transition for eight bars are repeated twice.

2.2.2. Chord Progression Generation

Next, chord progression is generated randomly. In the proposed system, only diatonic chords (I ~ VI) are used. Based on the rule shown in Table 1, chord progression in each motif is generated randomly. Last chord is selected from tonic chord.

2.3. Rhythm Sequence Generation

The rhythm sequence generation process has two phases; (1) generation of rhythm for basic motif and (2) generation of rhythm for derivation motif.
Table 1: Chord Progression Rule

<table>
<thead>
<tr>
<th>From</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
<th>VI</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>II</td>
<td>NG</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>III</td>
<td>NG</td>
<td>NG</td>
<td>OK</td>
<td>NG</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>IV</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
<td>NG</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>V</td>
<td>OK</td>
<td>NG</td>
<td>NG</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
</tr>
<tr>
<td>VI</td>
<td>NG</td>
<td>OK</td>
<td>NG</td>
<td>OK</td>
<td>OK</td>
<td>OK</td>
</tr>
</tbody>
</table>

(1) Generation of Rhythm for Basic Motif  
Rhythm for basis motif are generated randomly. If a note which is used mainly is quarter note and eighth note, quarter note or eighth note + eighth note are selected randomly. If a note which is used mainly is eighth note or sixteenth note, eighth note or sixteenth note + sixteenth note are selected randomly. And then, rhythm which including rest (for example, eighth note + eighth rest) are assigned to the last of motif if needed. Moreover, triplet and dotted note is assigned if needed.

(2) Generation of Rhythm for Derivation Motif  
Rhythm for derivation motif are generated based on derivation. In each motif, one block (one beat) is selected randomly, different rhythm pattern is assigned. For example, quarter note is replaced by eighth note + eighth note.

2.4. Pitch Assignment by Genetic Algorithm

Based generated chord progression and rhythm sequence, pitch is assigned by genetic algorithm.

2.4.1. Flow of Pitch Assignment

Step 1 : Initial Population Generation  
In the proposed system, initial individuals are generated randomly.

Step 2 : Fitness Calculation  
The fitness of each individual is calculated. In the proposed system, the fitness of the gene is calculated based on (1) pitch difference between two sounds, (2) successive non-harmonic tones, (3) successive disjunct motion over four degree and (4) last tone.

Step 3 : Selection  
Based on fitness calculated in Step 2, individuals used in Step 4 (crossover) are selected by the roulette selection and the elite preserve strategy.

Step 4 : Crossover  
New individuals are generated from the parents which are selected in Step 3 by the multi-point crossover.

Step 5 : Mutation  
In order to maintain genetic diversity, the mutation is carried out.

Step 6 : Repeat  
Steps 2 – 5 are repeated $T_{max}$ times.

Figure 1: Gene and Corresponding Melody.

Table 2: Gene Expression (Basic Motif).

<table>
<thead>
<tr>
<th>Chord</th>
<th>Candidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>C</td>
</tr>
<tr>
<td>II</td>
<td>F</td>
</tr>
<tr>
<td>III</td>
<td>G</td>
</tr>
</tbody>
</table>

2.4.2. Gene Expression  
In the proposed system, pitches for basic motif are expressed as gene. For derivation motif, variation rule are expressed as gene. For basic motif, each part in a gene takes 0–59, and the value divided by the number of pitch candidates means the pitch. For derivation motif, each part in a gene takes rule index.

Figure 1 shows an example of gene and corresponding melody. In Fig.1, red rectangle shows non-harmonic tone selected randomly. Table 2 shows a decision process from gene expression to corresponding melody.

2.4.3. Fitness

(1) Pitch Difference between Two Sounds  
In the proposed system, the fitness on pitch difference between two sounds of the gene $p$, $V_{TD}(p)$ is calculated by

$$V_{TD}(p) = \frac{1}{N^m} \sum_{m=1}^{N^m} \left(1 - D_{JS}(P_i^{m(p)} || P_i^{m(i)}) \right)$$  

(1)

where $N^m$ is the number of motifs. $D_{JS}(P_i^{m(p)} || P_i^{m(i)})$ is JS (Jensen–Shannon) divergence between pitch difference in the motif $m$ given by gene $P_i^{m(p)}$ and pitch difference distribution given by user $P_i^{m(i)}$, and it is given by

$$D_{JS}(P_i^{m(p)} || P_i^{m(i)}) = \frac{1}{2} D_{KL}(P_i^{m(p)} || P_i^{m(p,i)}) + \frac{1}{2} D_{KL}(P_i^{m(i)} || P_i^{m(p,i)})$$  

(2)
where $P_{m(p)}$ is the average of $P_{m(p)}$ and $P_{m(i)}$ and it is given by

$$P_{m(p;i)} = \frac{1}{2} (P_{m(p)} + P_{m(i)}) \quad (3)$$

$D_{KL}(P_{m(p)} || P_{m(p;i)})$ and $D_{KL}(P_{m(i)} || P_{m(p;i)})$ are KL (Kullback–Leibler) divergence of $P_{m(p)}$ and $P_{m(p;i)}$ and $P_{m(i)}$ and $P_{m(p;i)}$, and they are given by

$$D_{KL}(P_{m(p)} || P_{m(p;i)}) = \sum_{j=1}^{N^0} \left( P_{m(p)} \log_2 \frac{p_{m(p)}}{p_{m(p;i)}} \right) \quad (4)$$

$$D_{KL}(P_{m(i)} || P_{m(p;i)}) = \sum_{j=1}^{N^0} \left( P_{m(i)} \log_2 \frac{p_{m(i)}}{p_{m(p;i)}} \right) \quad (5)$$

If $P_{m(p)} = 0$ or $P_{m(i)} = 0$, they are given by

$$P_{m(p)} \log_2 \frac{p_{m(p)}}{p_{m(p;i)}} = 0 \quad (6)$$

If $P_{m(i)} = 0$ or $P_{m(p;i)} = 0$,

$$P_{m(i)} \log_2 \frac{p_{m(i)}}{p_{m(p;i)}} = 0 \quad (7)$$

In Eqs.(4), (5), $N^0$ is the number of categories of pitch difference, in the proposed system, $j = 1$ means 1st, $j = 2$ means 2nd, $j = 3$ means 3rd or 4th, $j = 4$ means 5th, $j = 5$ means over 5th, and so $N^0 = 5$.

JS divergence becomes 0 when two distributions are same.

(2) Successive Non-Harmonic Tones

The fitness on successive non-harmonic tones of the gene $p$, $V_{NC}(p)$ is given by

$$V_{NC}(p) = \frac{1}{N^m} \sum_{m=1}^{N^m} f_{NC}(N_{NC}^{m}(p)) \quad (8)$$

$$f_{NC}(u) = \begin{cases} 1 & (u = 0) \\ 0 & \text{(otherwise)} \end{cases} \quad (9)$$

where $N_{NC}^{m}(p)$ is the number of successive non-harmonic tones in the motif $m$ of melody given by the gene $p$.

2.4.4. (3) Successive Disjunct Motion over Fourth

The fitness on successive disjunct motion over 4th of the gene $p$, $V_{SK}(p)$ is given by

$$V_{SK}(p) = \frac{1}{N^m} \sum_{m=1}^{N^m} f_{SK}(N_{SK}^{m}(p)) \quad (10)$$

$$f_{SK}(u) = \begin{cases} 1 & (u = 0) \\ 0 & \text{(otherwise)} \end{cases} \quad (11)$$

where $N_{SK}^{m}(p)$ is the number of successive disjunct motions over 4th in the motif $m$ of the melody given by the gene $p$.

(4) Last Tone

The fitness on last tone of the gene $p$, $V_{LT}(p)$ is calculated by

$$V_{LT}(p) = \begin{cases} 1 & \text{(if last sound is root)} \\ 0.5 & \text{(if last sound is non-harmonictone)} \\ 0 & \text{(if last sound is non-harmonictone)} \end{cases} \quad (12)$$

(5) Total Fitness

Total fitness is calculated by

$$V(p) = V_{NC}(p) + 0.6 \times 10^{-10} \times \exp(V_{TD}(p) \times 23.65) + V_{SK}(p) + V_{LT}(p) \quad (13)$$

3. Computer Experiment Results

Here, we show the computer experiment results to demonstrate the effectiveness of the proposed system. Figure 2 shows examples of generated melodies. Figure 3 shows the fitness transition.
4. Conclusions

In this paper, we have proposed the automatic melody generation system considering chord progression. In the proposed system, chord progression and rhythm sequence are generated randomly, and the pitch is assigned to each note using genetic algorithm. We carried out a series of computer experiments, and we confirmed that melodies can be generated by the proposed system.

References


A Study on Satisfaction Method of Constraints by Approximating Constraints

Ooe Ryo† and Yoshikawa Tomohiro‡
†Graduate School of Engineering, Nagoya University
Furo, Chikusa, Nagoya, Aichi 464-8603, Japan
‡Graduate School of Engineering, Nagoya University
Furo, Chikusa, Nagoya, Aichi 464-8603, Japan
Email: ooe@cmplx.cse.nagoya-u.ac.jp, yoshikawa@cse.nagoya-u.ac.jp

Abstract—Genetic Algorithm (GA) is actively applied to real-world problems. Most of these real world problems are constrained optimization problems which optimize objective functions while satisfying constraints. In constrained optimize problems, the constraints need to be determined whether each of them is satisfied or not, which is regarded as a two-class classification. Therefore, the performance of GA for constrained optimization problems can be improved by classifiers as the approximation of constraints. In this paper, the approximation method of objective functions and constraints using Multiple Regression Analysis was applied to five benchmark problems as an introduction to the classification of constraints and studied the improvement of the performance of search in constrained optimization problems.

1. Introduction

Genetic Algorithm (GA) is actively applied to real-world problems such as the design of a front nose of N700, the design of aircraft wing of MRJ, and Nurse Scheduling Problem[2][3][4]. Most of these real world problems are constrained optimization problems which optimize objective functions while satisfying constraints. In the conventional GA researches, it has been reported that the performance of search can be improved by approximation of objective functions[5][6]. However, the approximation of fitness values of objective functions requires high approximation accuracy and it takes many individuals and actual evaluations as the information of approximation. Therefore, it is difficult to apply it, or it does not show high performance in real-world problems. On the other hand, constraints need to be determined whether each of them is satisfied or not, which is regarded as a two-class classification. And usually all of the constraints have to be satisfied. This paper studies the improvement of the performance of search in constrained optimization problems using Multiple Regression Analysis[1] as an introduction to approximation of constraints.

2. Proposed method

In GA, actual evaluation is generally required to acquire the fitness value(s) of objective function(s) and whether it violates constraint(s) for each generated offspring. Thus, offsprings which have worse fitness value(s) of objective function(s) than those of their parents or violate constraint(s) can be generated and that leads a waste of search. Therefore, in the proposed method, only offsprings which have better fitness value(s) than their parents and do not violate any constraints are generated based on Multiple Regression Analysis (MRA) and actual evaluation are done for them to improve the performance of search. In the proposed method, the crossover of same parents is repeated until the offsprings described above are generated or the number of crossover reaches the upper limit. The procedures of usual GA and the proposed method are shown in Fig.1.

3. Experiment

In this paper, the conventional method and the proposed method were applied to five benchmark problems (CF1 to CF5) which were employed for Multi-objective Optimization Test Instances for the CEC 2009 Special Session and Competition[7]. All of these benchmark problems, CF1 to CF5, are constrained optimization problems having one constraint and two objective functions. The details of these benchmark problems are shown in Table 1. The number of evaluations was 300,000 which was the condition used for the competition. In this experiment, the population size was 300 and the search was ended at 1,000th generation. In the proposed method, the upper limit of the number of crossover by same parents was 100 and the regression equations for the objective functions and constraint were calculated by newly obtained individuals at each generation.

The individuals satisfying the constraint for each benchmark problem are shown in Fig.2 to Fig.6. In all benchmark problems, the performance of the proposed method which uses MRA to approximate the objective functions and constraint was worse than that of the conventional GA. It is supposed that the worse performance of the proposed method was caused by the insufficiency accuracy of the approximation by MRA. The actual evaluated fitness values and the approximated values of the objective functions and constraint are shown in Fig.7 to Fig.11. The gradational color of each point in the figures is correspond to the gen-
eration, which changes from red (first generation) to blue (last generation). These figures show that the accuracy of the approximation was lower especially in early generations and it made the performance of the search worse. Because the more the search was conducted, individuals were more converged, it is thought that the accuracy of the approximation became better in later generations. However, remarkable inclinations of this accuracy improvement were not observed in these experiments. It cannot be denied that is the limit of a linear approximation method. In addition, as shown in Fig.7, extremely small approximated values by MRA was observed in CF1. It is supposed that was caused by the multicollinearity.

4. Conclusion

This paper studied the improvement of the performance of search in constrained optimization problems using Multiple Regression Analysis as an introduction to approximation of constraints. The proposed method which approximates objective function and constraints by MRA and generates offsprings by crossover considering them was applied to five constrained optimization problems and compared with the conventional GA. As the result, the performance of the proposed method was worse than that of the conventional GA. It was supposed that the low performance was caused by the insufficiency accuracy of the approximation by MRA especially in early generations. We will study on the approximation method for objective functions and constraints considering the convergence of individuals. We will also study on nonlinear approximation methods for the approximation.

References


<table>
<thead>
<tr>
<th>Table 1: Benchmark problems</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CF1</strong></td>
</tr>
<tr>
<td>$f(x) = x_1 + \frac{1}{2} \sum_{j=3}^{10} (x_1 - x_j)^2$</td>
</tr>
<tr>
<td>$f(x) = 1 - x_1 + \frac{1}{2} \sum_{j=3}^{10} (x_1 - x_j)^2$</td>
</tr>
<tr>
<td>$\mathbf{N} = 10$</td>
</tr>
<tr>
<td><strong>CF2</strong></td>
</tr>
<tr>
<td>$f(x) = 1 - \sqrt{x_1} + \frac{1}{2} \sum_{j=3}^{10} (x_1 - x_j)^2$</td>
</tr>
<tr>
<td>$\mathbf{N} = 10$</td>
</tr>
<tr>
<td><strong>CF3</strong></td>
</tr>
<tr>
<td>$f(x) = 1 - x_1 - \sqrt{x_1} + \frac{1}{2} \sum_{j=3}^{10} (x_1 - x_j)^2$</td>
</tr>
<tr>
<td>$\mathbf{N} = 10$</td>
</tr>
<tr>
<td><strong>CF4</strong></td>
</tr>
<tr>
<td>$f(x) = \sqrt{x_1} + \frac{1}{2} \sum_{j=3}^{10} (x_1 - x_j)^2$</td>
</tr>
<tr>
<td>$\mathbf{N} = 10$</td>
</tr>
<tr>
<td><strong>CF5</strong></td>
</tr>
<tr>
<td>$f(x) = 1 - x_1 - \sqrt{x_1} + \frac{1}{2} \sum_{j=3}^{10} (x_1 - x_j)^2$</td>
</tr>
<tr>
<td>$\mathbf{N} = 10$</td>
</tr>
</tbody>
</table>
Figure 1: Procedure of GA

Figure 2: Result of CF1
(a) Conventional GA  (b) Proposed method

Figure 3: Result of CF2
(a) Conventional GA  (b) Proposed method

Figure 4: Result of CF3
(a) Conventional GA  (b) Proposed method

Figure 5: Result of CF4
(a) Conventional GA  (b) Proposed method

Figure 6: Result of CF5
(a) Conventional GA  (b) Proposed method

Figure 7: Accuracy of CF1
(a) Objective function $f_1$  (b) Objective function $f_2$  (c) Constraint
A proposal of a new approach for strengthening the search ability of EMO algorithm -SPLASH-

Hibiki Samonji† and Shinya Watanabe†

†The Graduate School of Engineering, Muroran Institute of Technology
27–1 Mizumoto, Muroran, Hokkaido 050-8585 Japan
Email: 15043026@mmm.muroran-it.ac.jp, sin@csse.muroran-it.ac.jp

Abstract—In this paper, a new local search method using search history in evolutionary multi-criterion optimization (EMO) is proposed. This approach was designed by two opposite mechanisms (escaping from local optima and convergence search) and assumes to incorporate these into an usual EMO algorithm for strengthening its search ability. The main feature of this approach is to perform a high efficient search by changing these mechanisms according to the search condition. If the search situation seems to be stagnated, escape mechanism would be applied for shifting search point from this stagnated condition point to another points and breaking this stagnation. On the other hand, if it observes no sign of the improvement of solutions after repeating escape mechanism in a certain number of times, this approach judges that the solutions are near global optima and convergence mechanism is applied to improve their qualities by intensive local search. In this paper, this approach is called “a escaping from local optima and convergence mechanisms based on search history - SPLASH -”.

Experimental results show that the effectiveness of the proposed mechanisms was verified through investigating the influence by the presence or absence of the proposed each mechanism.

1. Introduction

In recent years, evolutionary multi-criterion optimization (EMO) is one of the most active research areas and various kinds of multi-objective evolutionary algorithm (MOEA) was proposed. Additionally, the combination MOEA and local search (LS) approach has been studied and presented its high search performance. These approaches are called multi-objective memetic algorithm (MOMA) and were widely proposed as typified by multi-objective genetic local search (MOGLS)[1].

In general, these approaches use a neighborhood search or a kind of gradient search and usually need high computational costs. Also, there has been some approaches for escaping from local optima in single-objective optimization[2, 3], but very few studies in multi-objective optimization.

Therefore, we proposed a new local search method using a search history in EMO, named SPLASH[4].

SPLASH is consists of escaping from local optima and convergence search mechanism. Escape mechanism estimates unexplored regions using whole search history information and tries to escape from local optima. On the other hand, convergence mechanism estimates prospective regions using a part of search history information and tries to improve a quality of solution.

The computational cost of each mechanism are same as that of genetic operation. thus, if we apply SPLASH instead of genetic operation, the computational cost in each generation is no difference to that of normal case (not using our mechanisms).

2. SPLASH

Our approach, named SPLASH, is consists of two mechanisms: escape mechanism and convergence mechanism. These mechanisms using search history information would be expected to strengthen the search ability of MOEA. In this section, we explain how to store the search information into a search history, then details of these mechanisms of SPLASH.

2.1. How to Store Search History

Since the amount of memory in a computer simulation must be limited, every search history information cannot be stored directly. Therefore, our approach uses the discretization of the search history information when to store these information into memory. This discretization is designed by reference to the concept of long term memory[4].

The concept of this storing approach is described in Figure 1. In Figure 1, feasible region in each variables is [0,1] and the memory of search history is discretized into three parts; [0,0.33), [0.33,0.67) and [0.67,1.0].

In Figure 1, the memory of search history is represented by a matrix and row of this matrix means the variable value and column presents the discretized range of each variable value.

2.2. Escape Mechanism

This mechanism tries to break through a search stagnating condition in design space using search history information. In order to detect a stagnating condition, this mechanism uses a stagnation parameter ki (i = 1,.,,N). Each convergence mechanisms based on search history[5].
individual (each sub-problem in MOEA/D) has own this parameter and increments this parameter by one \(k_i=k_i+1\) if the individual \(x'\) is not update after a selection operator. On the other hand, when the individual \(x'\) is updated, the parameter \(k_i\) is assigned to zero \(k_i=0\). And then, if \(k_i\) is over the pre-defined threshold \(K (k_i > K)\) at the beginning of next generation, escape mechanism would be performed instead of normal genetic operators for creating new solutions (such as crossover and mutation).

The flow of this mechanism is as follows. In here, \(M\) represents the memory matrix presented in 1. This matrix \(M\) is used in escape mechanism and \(m_{ij}(i = 1, \ldots, D \text{ and } j = 1, \ldots, n)\) is a element of \(M\). Also, \(D\) is the number of rows.

We defined memory \(M\) each sub problems and count neighborhood solutions.

**Step1. Selecting update range**
Select candidate solutions for the seeds of new solution using \(rand\), where \(rand\) is uniformly random number from \([0, 1]\). In here, we used two different ways of choosing candidate depending the value of \(rand\) like below equation.

\[
P = \begin{cases} 
B(i) & \text{if } \text{rand} < \delta \\
[1, \ldots, N] & \text{otherwise} 
\end{cases}
\]

The above function returns the suffix numbers of candidate solutions and vector \(P\) stores the suffix information of candidate solutions. \(B(i)\) is the function that returns around the value of input “\(i\)”. On the other hand, \([1, \ldots, N]\) return a random integer number from 1 to \(N\). This selection of vector \(P\) has a role of defining the range of updating for sub-problem in MOEA/D.

**Step2. Selecting variables for changing**
Randomly select some variables values in \(x\).

**Step3. Inverting the value of memory \(M\)**
In order to invert the value of memory \(M\), each element of \(M\) is replaced by \(M = m_{ij}^{\text{new}} - m_{ij}(i = 1, \ldots, D \text{ and } j = 1, \ldots, n)\). In here, \(m_{ij}^{\text{new}}\) means the biggest value in \(j\) rows in \(M\).

**Step4. Deciding the value of variant**
Changing the value of the selected variables in Step2. In order to decide the changing value, it needs to decide the discretized range of the selected variables. Here a roulette selection approach for each rows in \(M\) is applied to select the discretized range. Through this roulette selection, the changing value \(y\) is randomly generated within this discretized range.

**Step5. Generate new solution**
A new solution \(x'\) is generated by replacing the values of the selecting variables of \(x\) with \(y\) generated by Step4.

**Step6. Update**
Our updating follows the solution updating mechanism of MOEA/D-DE[5]. The steps of updating solutions in MOEA/D-DE are follows. In the following step, counter parameter \(c\) is used for calculating the number of updating.

1) If \(c = n\), or \(P\) is empty, finish. Otherwise, randomly pick an index \(p\) from \(P\).
2) If \(g(x'[p], z) \leq g(x[p], z)\), then \(x'[p] = x'\) and \(c = c + 1\).
3) Delete \(p\) from \(P\) and go to 1).

**2.3. Convergence Mechanism**
This mechanism performs intensive local search around prospective areas in design space using search history information. Basic concept of this mechanism is the same as that of escape mechanism, but the role of these mechanisms are opposite. And there are two different points in terms of memory index. One point is that memory index consists of the information of near the current solution. That is, the memory used in this mechanism is same as that of escape mechanism, but focus on a part of this memory. And the second one is that the element value of memory matrix is not inverted like Step3 of escape mechanism.

The procedure of this mechanism is almost same as that of escape mechanism, but the above-mentioned search memory is quite different.

**3. Experimental Results**
In these experiments, firstly we compared two algorithms (original MOEA/D-DE[5] and MOEA/D-DE with SPLASH) to investigate the influence of SPLASH using WFG test suites. Secondly, we investigated the influence of each mechanisms in SPLASH through analyzing a transition of the execution ratio between two mechanisms of SPLASH and genetic operation.
3.1. Benchmark Problems

As benchmark problems, we used WFG test suites[6]. WFG consists of nine benchmark instances, WFG1-WFG9. The details of WFG are shown in [6]. In these experiments, we set the number of variables \( n = 20 \), the number of position parameters \( k = 2(M - 1) \) and the number of distance parameters \( l = n - k \), where \( M \) is number of objectives. However, only WFG3 has a degenerate Pareto front in the case of over 3 objectives. Since there is a part of non-degenerate Pareto front in WFG3, we used modified WFG3 proposed in [7].

3.2. Parameters

In these experiments, stopping criterion was 100000 function evaluations and average hypervolume value of 50 runs is used as a measure of obtained solutions. The setting of other parameter were as follows.

**MOEA/D-DE**
- decomposition parameter \( H = 199, 6, 4(m = 2, 5, 8) \)
- population size \( N = 200, 210, 330(m = 2, 5, 8) \)
- neighborhood size \( T = N/10 \)
- probability of mating/update in neighborhood \( \delta = 0.9 \)
- the maximum of individuals to update \( n_r = 2 \)
- scalarizing function Tchbycheff
- genetic operators
  - crossover rate \( CR = 1.0 \)
  - scaling parameter in DE operator \( F = 0.5 \)
  - mutation rate \( MR = 1/n \)
  - distribution index in polynomial mutation \( \eta = 20 \)
- SPLASH
  - the number of discretization of memory \( D = 25 \)
  - the number of rows in convergence mechanism \( D_{conv} = 5 \)
  - stagnation count to apply escape mechanism \( K = 5 \)

3.3. Metrics

The search performance of algorithms was evaluated by Hypervolume(HV)[8]. HV calculates the \( m \)-dimensional volume that obtained solutions dominate in objective space. High HV value shows good solutions in convergence, diversity and uniformity. In these experiments, we set reference point \( r = (3, 5, 7, 11, 13, 15, 17, 19) \).

3.4. Search Performance in WFG

The results are shown in Table 1 to 3. Our approach performed higher HV values than those of MOEA/D-DE in WFG4 and WFG9, but HV of WFG4 was more higher than that of WFG9. The main reasons could be thought that WFG4 and WFG9 have multi-modality and WFG9 has a parameter dependence. WFG7 and WFG8 also have parameter dependence. Therefore, our approach was not so good in the case of 2 objectives. WFG2 and WFG6 are unimodal problem and the results of these problems showed similar to those of WFG7 and WFG8. Additionally, our approach was good results in the case of many-objective problem.

As a consequence, our approach was efficient for multimodal problems and many-objective problems. On the other hand, our approach was good for unimodal problems and parameter depending problems.

3.5. Process of Search

Table 4 shows HV and stagnation count in WFG4. In this table, SPLASH indicates MOEA/D-DE with SPLASH, escape means MOEA/D-DE with only escape mecanism and convergence means MOEA/D-DE with only convergence mecanism.

Stagnation count is the number of individuals that are not updated in each generations and the maximum value is 100000. The highest HV value is performed by proposed and lowest is performed by original MOEA/D-DE and the highest value of stagnation count is also performed by original MOEA/D-DE.

A transition of the execution ratio are shown in Figure 2 to 4. In these figures, vertical axis represents each individual and horizontal axis is generations. Figure 2 shows escape mechanism is effective in middle part of the search and Figure 3 shows convergence mechanism is conducted in last part of the search.

4. Conclusions

In this paper, a new local search method using search history in EMO, SPLASH was proposed. Experimental results showed that the effectiveness of the proposed mechanisms was verified through investigating the influence with or without the proposed mechanisms. Escape mechanism is effective in middle part of the search and convergence mechanism is in last part of the search. In future work, we will research about dynamic parameter settings and how to store search history can be effective in problem having parameter dependence.

References


Table 1: HV of 2 objectives (×10^1)

<table>
<thead>
<tr>
<th>MOEA/D-DE</th>
<th>SPLASH</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFG1</td>
<td>0.570</td>
</tr>
<tr>
<td>WFG2</td>
<td>1.143</td>
</tr>
<tr>
<td>WFG3</td>
<td>1.095</td>
</tr>
<tr>
<td>WFG4</td>
<td>0.844</td>
</tr>
<tr>
<td>WFG5</td>
<td>0.815</td>
</tr>
<tr>
<td>WFG6</td>
<td>0.812</td>
</tr>
<tr>
<td>WFG7</td>
<td>0.868</td>
</tr>
<tr>
<td>WFG8</td>
<td>0.813</td>
</tr>
<tr>
<td>WFG9</td>
<td>0.832</td>
</tr>
</tbody>
</table>

Table 2: HV of 5 objectives (×10^4)

<table>
<thead>
<tr>
<th>MOEA/D-DE</th>
<th>SPLASH</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFG1</td>
<td>0.400</td>
</tr>
<tr>
<td>WFG2</td>
<td>1.001</td>
</tr>
<tr>
<td>WFG3</td>
<td>0.677</td>
</tr>
<tr>
<td>WFG4</td>
<td>0.653</td>
</tr>
<tr>
<td>WFG5</td>
<td>0.651</td>
</tr>
<tr>
<td>WFG6</td>
<td>0.697</td>
</tr>
<tr>
<td>WFG7</td>
<td>0.687</td>
</tr>
<tr>
<td>WFG8</td>
<td>0.525</td>
</tr>
<tr>
<td>WFG9</td>
<td>0.555</td>
</tr>
</tbody>
</table>

Table 3: HV of 8 objectives (×10^7)

<table>
<thead>
<tr>
<th>MOEA/D-DE</th>
<th>SPLASH</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFG1</td>
<td>1.570</td>
</tr>
<tr>
<td>WFG2</td>
<td>3.387</td>
</tr>
<tr>
<td>WFG3</td>
<td>2.084</td>
</tr>
<tr>
<td>WFG4</td>
<td>1.586</td>
</tr>
<tr>
<td>WFG5</td>
<td>1.609</td>
</tr>
<tr>
<td>WFG6</td>
<td>1.914</td>
</tr>
<tr>
<td>WFG7</td>
<td>1.755</td>
</tr>
<tr>
<td>WFG8</td>
<td>1.262</td>
</tr>
<tr>
<td>WFG9</td>
<td>1.464</td>
</tr>
</tbody>
</table>

Table 4: Result of WFG4

<table>
<thead>
<tr>
<th>MOEA/D-DE</th>
<th>SPLASH</th>
<th>escape</th>
<th>convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>WFG4</td>
<td>0.844</td>
<td>0.864</td>
<td>0.852</td>
</tr>
</tbody>
</table>

Figure 3: A transition of the execution ratio in 500 generations (MOEA/D-DE with only convergence mechanism)

Figure 4: A transition of the execution ratio in 500 generations (MOEA/D-DE with SPLASH)


A proposal on a new efficient framework dedicated to large scale vehicle routing problems -FOCUS-

Masashi Ito† and Shinya Watanabe‡ and Kazutoshi Sakakibara††

†Department of Information and Electronic Engineering, Muroran Institute of Technology
27–1,Mizumoto-cho,Muroran,Hokkaido 050-8585,Japan
‡College of Information and Systems, Muroran Institute of Technology
27–1,Mizumoto-cho,Muroran,Hokkaido 050-8585,Japan
††Faculty of Engineering, Toyama Prefectural University
Imizu,Toyama 939-0398,Japan
Email: 15043007@mmm.muroran-it.ac.jp, sin@csse.muroran-it.ac.jp, sakakibara@pu-toyama.ac.jp

Abstract—A new framework dedicated to large scale vehicle routing problems (VRPs) are proposed. The main characters of this framework are based on the way of divide and conquer strategy and to be aimed for producing an effect search in large scale problems. This framework using divide and conquer strategy has two contradictory mechanisms; problem decomposition and gradual restoration of decomposed ones. The former one is compatible with divide part and the latter is conquer part. In addition, we incorporated a circular partitioning scheme and an intensive search scheme around borderlines of partitionings. In this paper, the proposed framework is called “a framework based on divide and conquer strategy for very large scale VRPs -FOCUS-”.

To investigate the effectiveness of FOCUS, some test problems of which features that are already-known were used and examined how these mechanisms of FOCUS works in search process.

1. Introduction

Vehicle Routing Problems (VRPs) which are called as delivery planning problems are well known as combinatorial optimization problems and have attracted a great deal of attention since 1970’s due to their wide applicability and economic importance[1].

Although the objective of most VRPs’ application is to minimize the total area distance, VRPs inherently have multi-objective aspects such as the number of vehicles or the degree of dispersion between the distances of each vehicle. Therefore, there have been many studies using evolutionary multi-criterion optimization (EMO) algorithm to optimize multi-objective VRPs [2, 3].

Recently, data size and problem size become larger scale according to technical advantages of storage performance and cloud technology. Since this trend cause a new formidable issue as the combinatorial explosion and the increased computational cost, previous approaches is difficult to obtain the solutions to fill required quality in real time.

Therefore, we proposed a new framework based on divide and conquer strategy for very large scale VRPs[4]. In here, abbreviated name of the proposed framework as FOCUS1. This paper presented a improved version of FOCUS and presented the efficiency of FOCUS through the numerical examples using some different instances.

2. The modification of FOCUS

The main feature of FOCUS is based on divide and conquer strategy, and FOCUS has two contradictory mechanisms; problem decomposition and gradual restoration of decomposed ones. The former one is compatible with divide part and the latter is conquer part. In the FOCUS, firstly, the proposed approach tries to divide the whole area of original problem into some small areas as many as the number of depots. And then each small area is subdivided into sub-areas until the total number of sub-areas reaches to the pre-defined number. After finishing this area segmentation processing, gradual integration would be performed until every sub-area are integrated into the one (the original area).

However, FOCUS was difficult to form a route striding over a borderline because of the influence of area segmentation. In order to overcome this shortage, we implemented a circular partitioning scheme and a intensive search scheme around borderlines of segmentations.

The following is the specific flow of the proposed approach, and the flowchart figure of this process is shown in Fig. 1.

2.1. Proposed method in the past (FOCUS)

In this proposed method in the past, we proposed a new EMO approach for very large scale VRPs. The proposed approach has area segmentation and gradual area integration mechanisms.

Area segmentation

1FOCUS is abbreviated name of “a framework based on divide and conquer strategy for very large scale VRPs”.

- 451 -
The important point of area segmentation in MDVRPs is to divide an area so that each sub-area belongs to only one depot. Since this restriction enable to skip the decision of which depots customers in sub-area should belong, the divided sub-area could be treated as one of small size CVRPs with single depot.

Therefore, our segmentation mechanism is based on a multistep segmentation approach, namely, our segmentation firstly tries to divide the whole area of original problem into some areas as many as the number of depots, and then subdivides each area into sub-areas. Specifically, first step segmentation is implemented by assigning each customers to the nearest depots and in second step each area is subdivided according to the midpoint of line connecting points of two customers; one customer is farthest removed from depot and another is farthest removed from this customer.

**Gradual area integration**

Area integration mechanism tries to restore the segmentation sub-areas to original area by integrating each sub-areas. The key points of area integration are the timing of integration and the choices of integration sub-areas.

In our mechanism, this timing is when a best incumbent solution remains unchanged for a certain predefined generation and the choices of integration is selected in the inverse order of dividing area. This means that the last divided sub-areas are firstly integrated and the first divided ones are lastly integrated.

And the main reason why we adopt ‘gradual’ integration is to reduce the magnitude between problem settings before and after applying our integration. If every sub-areas are integrated to one at a time, the magnitude of difference between before and after integrating sub-areas are so large that the obtained information is not able to effectively utilize subsequent search. Therefore our integration mechanism used to integrate partial sub-areas in stages and repeat this integration until all sub-areas are unified. We expected that this gradual approach leads to an efficiently search, because the small magnitude of problem change before and after would enable that the obtained information of prior problem is effective seeds in posterior search.

### 2.2. The modified parts

In order to decrease the influence of area segmentation, we incorporated two modifications; a circular partitioning scheme and a intensive search scheme around borderlines. The details of these schemes are shown as follows.

**Circular partitioning scheme**

Circular partitioning scheme is one of segmentation method that was proposed by Haimovitch-Rinnooy Kan[5]. This method is well known that is very effective for large scale problem.

**Intensive search around borderlines**

Gradual area integration of FOCUS searches while restoring sub-areas to original area size. However, when segmentation area size returned to original area size, formed routes are strongly influenced by area segmentation.

Therefore, when segmentation area size restore sub-areas to original area size by gradual area integration, we incorporated search mechanism near area segmentation lines. This search mechanism is known to create a route overlapping area segmentation lines.

The followins are the detail steps of this mechanism.

**Step 1:** Set the parameter $count$ representing the generation of using intensive search around borderlines mechanism $count = 0$.

**Step 2:** In the case of two areas, find a perpendicular line from a line between each depots and draw a perpendicular line through each depots and select route($t$) in random order between two perpendicular lines. In the case of over three areas, find gravity point from each depot and select route($t$) in random order from circle centering on gravity point.

**Step 3:** To select route on limiting route, applying the following two methods, after that, increment $count$:

- According to the order distance between customer of route $t$ and customer of route in limiting range.
- According to the order distance between gravity point of route $S$ and gravity point of route in limiting range.

**Step 4:** If $count \neq T$, go back to Step2. However, in the case of $count = T$, terminate this mechanism.

3. Numerical experiments

We used two test problems of Cordeau’s instances from VRP website:\(^2\) p08, pr06. In addition, we used large scale

\(^2\)VRP website

http://www.bernabe.dorronsoro.es/vrp/
problems (X-n1001-k43-D2 : XD2) that we originally created. The details of three test problems are shown in Table 2.

In this experiment, we investigated the characteristics and effectiveness of the proposed approach by comparing the improved method to without FOCUS mechanisms (normal method) in p08.

3.1. Results and Analysis

The transitions of the total travel distance $F_{\text{dist}}$ in p08 and XD2 are shown in from Fig. 2 to Fig. 4. And the distribution maps of the final results in three problems are shown in Fig. 5 respectively.

As shown in from Fig. 2 to Fig. 4, The improved methods have a better transition than the normal method in XD2 of large scale problem. However, the improved methods were comparable with the normal method in p08 of small scale problem.

In p08, we got the different result from the cases of only using circular partitioning scheme and the case of others. In other words, we considered that area segmentation had a bad effect on search in the case of using circular partitioning scheme. However, at the beginning of search, FOCUS with circular partitioning scheme could obtain better solutions than that of normal method (Fig. 3).

From these results, in the case of small scale problems, this segmentation scheme can work well on original problem environment. Therefore, when to use this scheme, it is necessary to restore sub-areas to a original area in the early stages.

Here, we compared the result of original FOCUS (Fig. 9) to those of others methods (Fig. 7 - Fig. 8). From these figures, in the case of original FOCUS, can not form a route overlapping area segmentation lines. However, by
Table 1: Used Parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>p08</th>
<th>pr06</th>
<th>XD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of population N</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>Archive size A</td>
<td>250</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>The number of area division P</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>The upper period of the search stagnation G_S</td>
<td>5000</td>
<td>10000</td>
<td>5000</td>
</tr>
<tr>
<td>The upper period of intensive search around borderlines T</td>
<td>50</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 2: Problem Instance.

<table>
<thead>
<tr>
<th>Problem</th>
<th>p08</th>
<th>pr06</th>
<th>XD2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The number of customers</td>
<td>249</td>
<td>288</td>
<td>1000</td>
</tr>
<tr>
<td>The number of depots</td>
<td>2</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

Figure 8: Actual route of the final solution obtained by the determinate search method using gravity’s spacing.

Intensive search around borderlines, we could mitigate the influence caused by intensive search around borderlines can succeed in decreasing, the influence of area segmentation. On the other hand, in comparison with the optimal route(Fig. 6), same route did not exist. Therefore, it doesn’t need to call intensive search around borderlines heavily.

4. Conclusions

In this paper, we proposed “a framework based on divide and conquer strategy for very large scale VRPs -FOCUS-”, and incorporated circular partitioning scheme and presented two new mechanisms; intensive search around borderlines for large scale vehicle routing problems (VRPs).

We investigated the effectiveness of the improved approach by comparison of its performance with that of original FOCUS method. In the numerical experiments, the incorporated methods have a better transition than the normal method in a large scale problem.

As future works, we want to implement mechanisms to avoid the influence of partial solutions in sub-area. Specifically, we would like to implement mechanisms of alternating between area segmentation and area integration without returning to the original area size.

Figure 9: Actual route of the using only FOCUS method in p08

References


Application of Artificial Bee Colony Algorithm to Maximum Power Point Tracking in Photovoltaic Systems

Hideto Sakamoto† and Toshimichi Saito†

†Department of Electronics and Electrical Engineering, Hosei University
3–7–2 Kajino-cho, Koganei, Tokyo 184-8584, Japan
Emails: hideto.sakamoto.9u@stu.hosei.ac.jp; tsaito@hosei.ac.jp

Abstract—This paper studies application of the artificial bee colony algorithm to maximum power point tracking in Photovoltaic systems. Depending on insolation and temperature, the voltage-power characteristic becomes a complex multi-model shape and the maximum power point becomes time-variant. In order to track the maximum power point, this paper presents an improved algorithm including flexible re-assignment of individuals. Performing basic numerical experiments, the algorithm efficiency is investigated. The results are compared with several existing algorithms.

1. Introduction

The artificial bee colony algorithm (ABC) is known as an effective search algorithm in the swarm intelligence [1] [2]. The ABC is inspired by gulping behavior of honeybees and consists of two sub-routine: many points search and probabilistic search. The ABC is simple in concept, is easy to implement, and has been applied to various engineering systems including power electronics, analog-to-digital converters, and digital filters.

This paper studies application of the ABC to the maximum power point tracking (MPPT) of voltage-power characteristics in photovoltaic systems. The photovoltaic systems have studied as important renewable energy supply systems. The MPPT is an important problem in efficient renewable energy supply and has been studied extensively [3] [4]. However, depending on insolation and temperature, the voltage versus power characteristics become complicated dynamic multi-peak shape. It is not easy to realize the MPPT in such complicated characteristics.

In order to approach the MPPT, this paper presents an improved version of the ABC. In the ABC, the best individual is preserved and the other ones are re-assigned depending on stagnation of the search. The re-assignment is able to track the dynamic MPP flexibly.

Performing basic numerical experiments, the algorithm efficiency is investigated. The results are compared with several existing methods including individual swarm optimizers (PSOs [5]).

2. Photovoltaic system and cost function

Figure 1 shows the photovoltaic system. We apply the ABC to the MPPT in this system. The voltage-current characteristic is described by

\[
i_j = f(v_j, S_j) = I_{ph} - I_s \left( \exp \left( \frac{q v_j}{k A T n_s} \right) - 1 \right)
\]

\[
I_{ph} = (I_{sc} + k(T_s - T_r)) \frac{S_j}{100}, \quad j = 1 - 3
\]

where \( I_{ph} \) is the photo-generated current and \( n_s \) is the reverse saturation current. \( A \) is the diode ideality factor, and \( T_r \) is the temperature of solar cell. For simplicity, \( T_s \) does not vary as time goes. \( S_j \) is insolation of solar cells. Figure 2 shows insolation signals defined by

\[
S_1 = 15 \cos \left( \frac{5 t}{16} + \frac{7 \pi}{6} \right) + 85
\]

\[
S_2 = 20 \cos \left( \frac{3 t}{16} + \frac{5 \pi}{6} \right) + 50
\]

\[
S_3 = 10 \cos \left( \frac{t}{16} + \frac{\pi}{2} \right) + 30
\]

Since the voltage-current characteristic is one-to-one for its, we describe the characteristic of each cell as a function of current and time:

\[
v_1 = g_1(i_1) = f^{-1}(v_1, S_1)
\]

\[
v_2 = g_2(i_2) = f^{-1}(v_2, S_2)
\]

\[
v_3 = g_3(i_3) = f^{-1}(v_3, S_3)
\]

The whole characteristic is given by

\[
v = G(i) = \begin{cases} 
  g_1(i_1) & (i_3 \leq i < i_1) \\
  g_2(i_2) & (i_2 \leq i < i_3) \\
  g_3(i_3) & (0 \leq i < i_2)
\end{cases}
\]

Figure 3 is shows each voltage-current characteristic and whole voltage-current characteristic. Using this function, we obtain the objective time-variant cost function of the voltage-power characteristic:

\[
F(v) = vi = vG^{-1}(v)
\]

Figure 4 is shows snapshots of the cost function. Our purpose is to track time-variant MPPs in this cost function.
3. Artificial Bee Colony algorithm

In order to define the ABC, we give several basic definitions. The problem is tracking MPPs of the cost function in Equation (5).

ABC has 3 steps in search. The first step is global search. This step updates all particles depending on themselves and reference particles. The second step is local search. This step updates only one individual depending on relative value probability. Last step is re-assign. This step is re-assigned by some particles when stop updating.

In the conrational ABC, all particles update at the same time. But, operating point is only one point at MPPT in real systems. So, it is difficult that all individuals update at the same time. This paper defines only one particles update at every sampling time $t = n\Delta t$, where $\Delta t = 1/M$ is the sampling interval, $M$ is the number of particles. Here we defined the algorithm.

**Step 1 Initialization**
Using the voltage $v(t)$, the individual positions are initialized.

$$x_n = v(\Delta t)$$

Each individual has a counter $T_n$. $T_n$ is initialized $T_n = 0$. If individuals do not update. $T_n$ is updated; $T_n \leftarrow T_n + 1$.

**Step 2 Global search**
According to Equation (7), a candidate individual $x_c$ is generated.

$$x_c = x_n + \phi(x_n - x_r)$$

where $x_r$ is the reference individual. Reference individual is chosen randomly. $\phi$ is a random number in $[-1, 1]$. According to Equation (8), an individual is updated.

$$x_n \leftarrow \begin{cases} x_c & (F(x_c) > F(x_n)) \\ x_n & (F(x_c) \leq F(x_n)) \end{cases}$$

**Step 3 Local search**
According to Equation (9), all individuals make relative value with probability $P_n$

$$P_n = \frac{F(x_n)}{\sum_{m=1}^{M} F(x_m)}$$

Using relative value probability $P_n$, one individual is chosen.

**Step 4** The individual chosen by Step 3 repeats Step 2 for $M$ times.

**Step 5 Replace**
If $T_n$ exceeds a threshold limit $T_{lim}$, $x_n$ is re-assigned into search area.

**Step 6 Update**
The best individual Chose in all individuals.
Step 7 Repeat
Let $n \leftarrow n + 1$, go to step 2, and repeat until $n \Delta t = t_{\text{max}}$, where $t_{\text{max}}$ is maximum time.

Replacing Step 5 with the following Step 5’, we obtain the RABC.

Step 5’ Replace
If $T_n$ exceeds a threshold limit $T_{\text{lim}}$ and $x_n$ is not best individual then $x_n$ is re-assigned into search area.

4. Experiments

We apply the Rule-changed ABC (RABC) to the cost function Equation (5). After trial-and-error, the parameters are selected: $\Delta t = 0.2, t_{\text{max}} = 30, M = 5, T_{\text{lim}} = M$.

Figure 5 shows snapshots in trace process. For $t > 0$ can search the MPPs. The individuals have not accumulated and have kept diversity. Figure 6 shows MPPT process of ABC and RABC. RABC and ABC can almost trace MPP. But the individuals is apart from the MPP for $10 < t < 20$. RABC returns the MPP for $t < 16$. It is efficiency that the condition of re-assigned changes. After 100 times trials, in order to quantify the trace performance of the MPP, we have used Equation (10).

$$P_{ef} = \frac{\sum_{k=1}^{100} P_k \text{error} [\%]}{\sum_{k=1}^{100} (P_{\text{ref}} - M\text{PP}_n)}$$

$$P_k \text{error} = \frac{P_k - M\text{PP}_n}{M\text{PP}_n}$$

$$P_k = \text{power of time } t$$

$$M\text{PP}_n = \text{MPP at time } t$$

Table 1 shows performance of the MPPT of RABC, ABC and SDLPSO. The $P_{ef}$ of RABC is compared with that of ABC and SDLPSO.

<table>
<thead>
<tr>
<th>Table 1: MPPT performance in three algorithms</th>
</tr>
</thead>
<tbody>
<tr>
<td>RABC</td>
</tr>
<tr>
<td>$P_{ave}$</td>
</tr>
</tbody>
</table>

Figure 5: Snapshots of trace process of RABC. (a) $t = 0$ [s]. (b) $t = 30$. (c) $t = 60$. (d) $t = 90$. (e) $t = 120$. (f) $t = 150$.

Figure 6: MPPT process.
5. Conclusion

The RABC is presented and is applied to the MPPT in this paper. The algorithm include flexible re-assignment of individuals and can be effective to track dynamics MPPTs. The algorithm performance is investigated in basic numerical experiments and has been compared with several existing algorithms.

Future problems include analysis of search process and optimization of algorithm parameters.

References


Multi-point search algorithm with rotation angle dependent on the best position

Kenta Kohinata† Takuya Kurihara† Takuya Shindo† Kenya Jinno†

†Nippon Institute of Technology
‡Gakuen-dai 4-1, Miyashiro-Saitama, Saitama, 345-8501, Japan
Email: e1112228@estu.nit.ac.jp, e1072215@estu.nit.ac.jp, t-shindo@nit.ac.jp, jinno@nit.ac.jp

1. INTRODUCTION

Optimization problems to search the most suitable value of the given evaluation function are fundamental and important problems. However, it is generally difficult to search a feasible solution of large-scale problems within real time. In order to solve such large scale problems quickly, various heuristic optimization solvers are proposed. The particle swarm optimization (abbr. PSO) [1][2] is one of such optimization solvers. The PSO can search a feasible solution quickly comparing with other heuristic optimization solvers. Each particle of the PSO has the best position information that gives the best evaluation value, and it shares in swarm.

Since the PSO contains stochastic factors, the rigorous analysis of the dynamics of the PSO is quite difficult[3]. To analyze the dynamics, we proposed a canonical deterministic PSO (abbr. CD-PSO)[4].

The coordinate system of CD-PSO is adapted a canonical coordinate system. Since CD-PSO is a deterministic system, the dynamics of the particle of CD-PSO can be clarified theoretically. The analysis results of CD-PSO indicate that the search range is shrunk with the update of the best location information. To overcome this situation, we proposed the method to keep the extent of the search region[5]. However, the method cannot search around the obtained best position. Also, the method cannot generate diversity of the solutions because CD-PSO is a deterministic system. The diversity is very important for heuristic optimization solvers. The diversity of the solution search performance of the deterministic system is poorer than the stochastic system. To overcome this situation, we consider a mechanism to generate the diversity for the deterministic system. The mechanism improves the solution search performance of the local search. In this article, the purpose of this study is to confirm the solution search performance. We will confirm the solution search performance by using the benchmark functions.

2. CD-PSO

The conventional PSO is described by the following equations.

\[
\begin{align*}
    v_i^{t+1} &= w v_i^t + c_1 r_1 (p_{best_i}^t - x_i^t) + c_2 r_2 (g_{best}^t - x_i^t) \\
    x_i^{t+1} &= x_i^t + v_i^{t+1}
\end{align*}
\]

where, \(v_i^t\) and \(x_i^t\) denote the velocity vector and the location vector of the \(i\)-th particle on the \(t\)-th iteration, respectively. \(p_{best_i}^t\) means the location that gives the personal best value of the evaluation function of the \(i\)-th particle until the \(t\)-th iteration. \(g_{best}^t\) means the location which gives the best value of the evaluation function on the \(t\)-th iteration in the swarm. \(w \geq 0\) is an inertia weight coefficient, \(c_1 \geq 0\), and \(c_2 \geq 0\) are acceleration coefficients, and \(r_1 \in [0,1]\) and \(r_2 \in [0,1]\) are two separately generated uniformly distributed random numbers.

To analyze the dynamics of the conventional PSO, the random coefficients have been omitted from the conventional PSO. We rewrite the best location information as follows.

\[
\begin{align*}
    p_i^t &= \frac{c_1 p_{best_i}^t + c_2 g_{best}^t}{c} \\
    c &= c_1 + c_2
\end{align*}
\]

We normalize the location information by \(p_i^t\). Without loss of generality, we consider one-dimensional case. In this case, Eq. (1) is transformed into the following matrix form:

\[
\begin{bmatrix}
    y_i^{t+1} \\
    y_{i-1}^{t+1}
\end{bmatrix} =
\begin{bmatrix}
    w & -c \\
    w & 1 - c
\end{bmatrix}
\begin{bmatrix}
    y_i^t \\
    y_{i-1}^t
\end{bmatrix},
\]

where, \(y_i^t = x_i^t - p_i^t\).

The behavior of the particle is governed by the eigenvalues of the matrix in Eq. (3). The eigenvalue \(\lambda\) is derived as follows.

\[
\lambda = \frac{(1 + w - c) \pm \sqrt{(1 + w - c)^2 - 4w}}{2}
\]
This system is a discrete-time system. Therefore, if the eigenvalues exist within the unit circle on the complex plane, the system is said to be stable. When the eigenvalues are complex conjugate numbers, the behavior of the trajectory of Eq. (3) in the phase space $v_i - y_i$ becomes a spiral motion. We have clarified that the system exhibits an excellent solution search performance when the particle exhibits the spiral motion in the phase space. In this case, the damping factor $\Delta$ and the rotation angle $\theta$ are given as follows.

$$\Delta = \sqrt{\text{Im}(\lambda)^2 + \text{Re}(\lambda)^2} = \sqrt{w},$$  \hfill (5)

$$\theta = \arctan \frac{\text{Im}(\lambda)}{\text{Re}(\lambda)} = \arctan \frac{\sqrt{4w - (1 + w - c)^2}}{(1 + w - c)}. $$  \hfill (6)

To clarify the effect of the eigenvalues, we derive a canonical deterministic PSO (abbr. CD-PSO).

$$\begin{bmatrix} \frac{y_{i+1}}{v_{i+1}} \\ \frac{v_{i+1}}{v_{i+1}} \end{bmatrix} = \begin{bmatrix} \delta & -\omega \\ \omega & \delta \end{bmatrix} \begin{bmatrix} \frac{y_i}{v_i} \\ \frac{v_i}{v_i} \end{bmatrix} $$  \hfill (7)

The damping factor $\Delta$ and the rotation angle $\theta$ of CD-PSO are derived as

$$\Delta = \sqrt{\delta^2 + \omega^2},$$  \hfill (8)

$$\theta = \arctan \frac{\omega}{\delta}. $$  \hfill (9)

By using these parameters, Eq. (7) is rewritten as follows.

$$\begin{bmatrix} \frac{y_{i+1}}{v_{i+1}} \\ \frac{v_{i+1}}{v_{i+1}} \end{bmatrix} = \Delta \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} \frac{y_i}{v_i} \\ \frac{v_i}{v_i} \end{bmatrix} $$  \hfill (10)

If the rotational angle is set by the golden angle, the search position becomes not overlapped[6][7]. The golden angle $\phi$ is defined as

$$\phi = 180(3 - \sqrt{5})[\text{deg}].$$  \hfill (11)

The particle of CD-PSO cannot escape from the local minimum if the particle traps the local minimum[6][7].

To improve the global search ability, we proposed a control method that set a lower limit of the rotation radius on the phase space[5]. When the rotation radius is smaller than a criterion value $R$, the proposed method resets the velocity of the particle as follows.

$$\begin{bmatrix} \frac{x_i}{v_i} \\ \frac{v_i}{v_i} \end{bmatrix} = \begin{bmatrix} p_i^j + R \cos \theta \\ R \sin \theta \end{bmatrix}, \text{ if } \sqrt{\frac{\left | p_i^j \right |^2}{x_i}} + \sqrt{\frac{\left | v_i \right |^2}{v_i}} < R. $$  \hfill (12)

We confirmed that the efficiency of the local search is reduced because the criterion of the rotation radius is set[5].

### 3. Proposed method

The time series of the search points of CD-PSO are shown in Fig. 1a. The search points are vibrated as that the center is the best location and the rotation radius is $R$.

Therefore, the location information is described as the following equation.

$$\begin{bmatrix} x_i^j \\ y_i^j \end{bmatrix} = R \cos(\theta_j) + p_i^j, $$  \hfill (13)

However, the searching motion is the local search performance is poor[5] . Since the rotation angle is a constant, the efficiency of the search in the vicinity of the best position is poor. Therefore, in order to improve efficiency of the local search ability, we propose a method which depends on the rotation angle at the best position . The proposed method is shown in the following

$$\begin{bmatrix} x_i^j \\ y_i^j \end{bmatrix} = R \cos(\theta_j) + p_i^j, $$  \hfill (14)

$$\theta_j = \theta_{j-1} + \left( \frac{|p_i^j - x_i^j|}{R} \right). $$  \hfill (15)

The step width of the particle of the proposed method varies with the rotation angle . If the amount of changing times the rotation angle is large, it performs a global search. On the other hand, when the rotation angle is small, the particle performs a local search . The rotation angle is determined by the best position and its location. Figure 1 shows a changing time series of the search position of the particle. If the rotation angle is a constant, it performs generally search.

However, the proposed method can confirm that we search for a lot of position neighborhood best than rotation angle is a constant.

In this case, the rotation angle is synchronized. This situation is problem. When the rotation angle is synchronized, the biased search performance is reduced . Since the best position of each particle of the proposed method is different, there is no bias without different rotation angle synchronization . Figure 2 shows the behavior of the particles on the two dimensional space. If the rotation angle is a constant, the particle does not search on the second quadrant and the fourth quadrant on the phase space . However, the proposed method can be confirmed that it has been all quadrant search.

### 4. Numerical experiment

We confirm the performance with the benchmark functions. The applying benchmark functions are represented in Table 1. We compare the numerical simulation results of CD-PSO , PSO and the proposed method . Table 2 shows the results of CD-PSO, PSO, and the proposed method. These results indicate that the solution search ability of the proposed method is excellent comparing with the CD-PSO. Sphere function and Rosenbrock function is a unimodal function, Rastrigin functions and Griewank function is a multimodal function In the case of multi-modal evaluation function, the search results of the proposed method correspond to the conventional PSO. However, in the case of unimodal evaluation functions, the conventional PSO performance is better than the proposed method. The reason is the proposed system is not converged.
5. Conclusion

We proposed the multi-point search algorithm which is based on the behavior of the CD-PSO. In addition, we confirmed the effectiveness of the proposed method by using benchmark problems. It is necessary to confirm which parameter is effected to the solution search performance. Furthermore, we would like to construct a novel model of the deterministic PSO whose solution search ability is better than the conventional stochastic PSO. These are our future problems.

Acknowledgment

This work was supported by JSPS KAKENHI Grant-in-Aid for Challenging Exploratory Research Number: 16K14271, Grant-in-Aid for Scientific Research(C) Number: 15K06077.

References


### Table 2: Simulation result

<table>
<thead>
<tr>
<th>Function</th>
<th>Manner</th>
<th>Mean</th>
<th>Best</th>
<th>Worst</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>CD-PSO</td>
<td>5.57e+03</td>
<td>1.89e+03</td>
<td>1.37e+04</td>
<td>2.76e+03</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>1.91e-42</td>
<td>1.19e-51</td>
<td>5.44e-41</td>
<td>9.75e-42</td>
</tr>
<tr>
<td></td>
<td>Propsed</td>
<td>5.95e-05</td>
<td>1.08e-05</td>
<td>2.39e-04</td>
<td>5.68e-05</td>
</tr>
<tr>
<td>Rastrigin</td>
<td>CD-PSO</td>
<td>6.31e+03</td>
<td>2.76e+03</td>
<td>1.36e+04</td>
<td>2.49e+03</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>1.28e+02</td>
<td>5.27e+01</td>
<td>2.04e+02</td>
<td>3.92e+01</td>
</tr>
<tr>
<td></td>
<td>Propsed</td>
<td>1.54e+02</td>
<td>5.91e+01</td>
<td>3.16e+02</td>
<td>5.37e+01</td>
</tr>
<tr>
<td>Rosenbrock</td>
<td>CD-PSO</td>
<td>2.72e+08</td>
<td>3.18e+07</td>
<td>1.01e+09</td>
<td>2.64e+08</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>3.13e+01</td>
<td>1.34e+00</td>
<td>2.05e+02</td>
<td>4.25e+01</td>
</tr>
<tr>
<td></td>
<td>Propsed</td>
<td>5.69e+02</td>
<td>1.03e+01</td>
<td>3.39e+03</td>
<td>8.79e+02</td>
</tr>
<tr>
<td>Griewank</td>
<td>CD-PSO</td>
<td>2.39e+00</td>
<td>1.36e+00</td>
<td>4.42e+00</td>
<td>6.99e+01</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>2.15e-02</td>
<td>0.00e+00</td>
<td>9.82e-02</td>
<td>2.27e-02</td>
</tr>
<tr>
<td></td>
<td>Propsed</td>
<td>2.93e-02</td>
<td>2.10e-08</td>
<td>7.68e-01</td>
<td>1.07e-01</td>
</tr>
<tr>
<td>Shifted Rastrigin</td>
<td>CD-PSO</td>
<td>4.30e+04</td>
<td>1.76e+02</td>
<td>6.17e+04</td>
<td>1.31e+04</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>1.74e+02</td>
<td>3.01e+01</td>
<td>5.24e+02</td>
<td>1.03e+02</td>
</tr>
<tr>
<td></td>
<td>Propsed</td>
<td>1.56e+02</td>
<td>5.54e+01</td>
<td>3.74e+02</td>
<td>5.91e+01</td>
</tr>
<tr>
<td>Shifted Rosenbrock</td>
<td>CD-PSO</td>
<td>1.38e+10</td>
<td>1.13e+03</td>
<td>2.81e+10</td>
<td>7.04e+09</td>
</tr>
<tr>
<td></td>
<td>PSO</td>
<td>1.85e+01</td>
<td>1.00e-02</td>
<td>7.60e+01</td>
<td>2.29e+01</td>
</tr>
<tr>
<td></td>
<td>Propsed</td>
<td>3.34e+02</td>
<td>1.16e+02</td>
<td>9.14e+02</td>
<td>1.82e+02</td>
</tr>
</tbody>
</table>


Figure 2: The behavior of the particles in on the 2-dimensional solution space
The Application of the ES Operation in Firefly Algorithm

Takuya Shindo† and Kenya Jin’no†

†Department of Electrical and Electronics Engineering, Nippon Institute of Technology
4–1 Gakuendai, Miyashiro, Minamisaitama, Saitama 345-8501, Japan
Email: t-shindo@nit.ac.jp, jinno@nit.ac.jp

Abstract—In recent years, many optimization methods using swarm intelligence have been studied by many researchers. Furthermore, evolutionary algorithms that are classified into nonlinear optimization, have been improved the obtained performance. In the framework of the multi-point search algorithm, we believe that the boundaries of these techniques are disappearing from the trend of recent years of research. In other words, the concept of evolutionary algorithm is considered to be applicable to other swarm intelligence algorithms. In this article, we consider applying the performance improvement method of evolutionary strategy to the firefly algorithm, which is one of the swarm intelligence algorithms.

1. Introduction

In recent years, many optimization methods using swarm intelligence has been studied by many researchers. For example, particle swarm optimization (PSO)[1][2], firefly algorithm (FA)[3][4], ant colony optimization (ACO)[5] and so on, there are many algorithms. These algorithms can be said to be the met strategy algorithm. The boundary between the evolution strategy algorithm for multi-point search typified genetic algorithms are becoming ambiguous. Since, in essence the swarm intelligence algorithm, Individual operation of evolutionary strategy algorithm is included. In PSO, methods of introducing evolution strategy algorithm have been proposed. In order to prevent premature convergence of search, method to use the elite strategy and mutation [6] and, it has been proposed method carry out the selection of the particle [7].

Therefore, by introducing the operation of explicitly evolutionary strategy algorithm to swarm intelligence algorithm, we will consider that the use of knowledge relating to the performance improvement of evolution strategy algorithm.

In this paper, in the firefly algorithm, which is one of the swarm intelligence algorithm, The behavior of individuals upon explicitly introducing operation of ES is considered.

2. Evolutionary Strategy

The framework of the general ES shown in Algorithm 1. In Initialization, the first generation individuals is generated. In addition, the the fitness of the generated individual is evaluated. After initialization, enter the evolutional loop.

A new individuals (offspring) from a parent population in recombination is generated. This operation is the presence of two types. First, characteristics of the parent individual is the dominant recombination that is inherited to offspring. The other one, an intermediate recombination, The mean value of a parent population is used as a feature of the offspring generations. Mutation operation generates a diversity of offspring. For individuals randomly selected, its characteristics are modified.

Selection is used to set the next generation of the population. This selection is performed based on such evaluation value.

3. Firefly algorithm

Firefly Algorithm (FA)[3][4] is also one of such metaheuristic algorithms. The FA is developed based on the characteristics of the blinking of natural firefly by Xin-She Yang et al. in 2007. The conventional FA is based on the following three rules[3][4]:

1. Fireflies are unisex so that one firefly will be attracted to other fireflies regardless of their sex.
2. The attractiveness is proportional to the brightness, and they both decrease as their distance increases. Thus for any two flashing fireflies, the less brighter one will move towards the brighter one. If there is no brighter one than a particular firefly, it will move randomly.
3. The brightness of a firefly is determined by the landscape of the objective function.

Algorithm 1 General framework of evolutionary algorithm

1: Initialization
2: repeat
3: Recombination
4: Mutation
5: Evaluation
6: Selection
7: until Termination criterion
Algorithm 2 FA algorithm

1: Initialize the fireflies
2: repeat
3: Determine a new \( \alpha_i \)
4: Evaluation
5: Determine the best solution
6: Vary attractiveness according Eq. (1)
7: \( t = t + 1 \)
8: until Termination criterion

The algorithm of the FA shown in Algorithm 2.

The dynamics of the FA is described by the following equation.

\[
x^t+1_i = x^t_i + \beta(x^t_j - x^t_i) + \alpha_t \epsilon^t_i,
\]

where \( x \) denotes the position of each firefly. \( \beta \) denotes the attractiveness of the \( i \)-th firefly receives from the \( j \)-th fireflies. \( \alpha_t \) represents the randomness varies with \( t \). \( \epsilon^t_i \) denotes a random number vector whose distribution is Gaussian, uniformed, and so on, on the \( t \)-th iteration.

Parameter \( \alpha_t \) controls the randomness. This parameter can be varied by the status update. For example, the parameter is defined as

\[
\alpha_t = \alpha_0 \delta^t, \quad 0 < \delta < 1,
\]

where, \( \alpha_0 \) is an initial random scale factor, \( \delta \) is the attenuation coefficient. In many applications, the parameter \( \delta \) is set as the interval between 0.95 and 0.97.

In the case where \( L \) denotes the width of the search range, the initial randomness parameter \( \alpha_0 \) is set as \( \alpha_0 = 0.01L \). On the other hand, if the randomness parameter is small, the step size of the random walk becomes small. Thus the system can do a local search. In order to get a balance between a global search and a local search, the randomness parameter is set as above. \( \epsilon \) is set a uniform random number \([-0.5, +0.5]\). If \( \alpha_0 \) is small, step of initial search is small. On the other hand, if \( \alpha_0 \) is large, in the initial search state of the random walk is observed.

Parameter \( \beta \) controls the attractiveness. In general, \( \beta_0 = 1 \) is applied. \( \gamma \) is coefficient representing the attenuation of the light intensity. \( \gamma \) is related to the scale of the \( L \). In general, it is possible to set a \( \gamma = 1/\sqrt{L} \).

Recombination and mutation in Sec. 2 in the FA can be regarded as have been carried out concurrently in Eq. 1. On the other hand, selection operation is not performed. In other words, it can be said that the next generation population inherits strongly the characteristics of a parent. Therefore, we propose a method of introducing the selection operation for the FA.

4. ES operations with FA

In ES, there are several methods of selection operation[8]. In proportional selection, selection probability is proportional to the evaluation value of the individual.

Algorithm 3 FA with selection operation

1: Initialize the fireflies
2: repeat
3: Determine a new \( \alpha \)
4: Evaluation
5: Determine the best solution
6: Selection operation
7: Vary attractiveness according Eq. (1)
8: \( t = t + 1 \)
9: until Termination criterion

Table 1: Parameter setting of FA

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha_0 )</td>
<td>0.01L</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>1.0</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>1/\sqrt{L}</td>
</tr>
<tr>
<td>( \delta )</td>
<td>0.97</td>
</tr>
</tbody>
</table>

rank-based selection is used as an absolute evaluation value. In the tournament selection, the \( q \) individual selected by a uniform sampling, selecting a further most the fitness was higher individual. According to sample size \( q > 1 \) in Tournament selection, selective pressure varies.

By performing these operations after the evaluation in the FA, Individual selection of is carried out.

5. Numerical simulations

In order to confirm the effect of selection operation have on solution search, using Sphere function to perform numerical experiments.

\[
f(x) = \sum_{d=1}^{D} x_d^2
\]

The search range is \([-10, 10]^D\). In addition, the number of trials 20 times, state update 5000 times, \( D = 20 \) dimensions, population was set to 40. Parameter settings of FA shown in the Table 1. Parameter settings shown in the table are recommended values that are utilized in many applications [4].

The results are shown in Fig. 1. The horizontal axis denotes the number of state update, and the vertical axis represents the evaluation value. Original FA is the most high solution search performance from the results. Result of varying selective pressure \( q \) are shown in Fig. 2. In the case where the selective pressure high, performance is hardly changed. In other words selection operation is not working effectively.

6. Conclusions

The selection operation of the ES was introduced to the FA, which is one of the swarm intelligence. When per-
forming selection operation, performance deteriorates. The cause is decrease of individuals of diversity according to selection operation. Therefore, we will consider the introduction also mutation and recombination.

References


Figure 2: Varying selective pressure $q$ in tournament selection
On a rotationally invariant of PSO

Yosuke Hariya, Takuya Shindo and Kenya Jin’no

Nippon Institute of Technology
Saitama 345-0826, Japan
Email: e1112350@estu.nit.ac.jp, t-shindo@nit.ac.jp, jinno@nit.ac.jp

Abstract—Particle swarm optimization (PSO) is a stochastic population-based algorithm that is designed for real-parameter optimization problems. PSO is a simple and powerful algorithm. However, the performance of PSO is degraded in the case of non-separable problems. In this article, we discuss rotationally invariant PSOs and its performance.

1. Introduction
Optimization problem is an important issue in various fields. Single-objective continuous optimization problem is a problem of finding a real-valued vector that minimizes an objective function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \). In particular, an optimization problem that the analytic form is not known is called as black-box optimization problem. Many stochastic population-based algorithms have been proposed to solve the black-box optimization problem.

Particle swarm optimization (PSO)\([1][2]\) is one of the stochastic population-based algorithms that is based on swarm intelligence. PSO is simple and powerful algorithm. However, its search performance is depended on the coordinate system of the objective function \([3][4][5]\). Such property is referred to as rotation variance, and this property is related to separability of the objective function \([4]\). In the black-box optimization problem, the algorithm whose search performance is affected by the property of the objective function is undesirable. To overcome this problem, several rotationally invariant PSO have been proposed \([6][7][8][9][10][11]\). In this article, we discuss the rotation variance of PSO and we introduce the typical rotationally invariant PSO.

Separability: If the function can be rewritten as \( f(x) = \sum_{i=1}^{n} f_i(x_i) \), the function \( f \) is said to be separable \([4]\). Namely, the function \( f \) corresponds to each dimension is independent.

In general, if the number of dimensions increases linearly, the volume of the search space increases exponentially. However, since the separable function can be rewritten as the sum of the 1-dimensional functions, the complexity of the problem increases linearly. Thus, the separable functions is said to be an easier problem than the non-separable function.

In almost cases, a separable function can be transformed into the non-separable function by rotation of the coordinate system. From this fact, the performance of the algorithm is depended on the separability of the objective function is referred to as rotation variance.

2. Particle Swarm Optimization

PSO has been proposed by Kenedy and Clerc \([1][2]\). Each particle contains three vectors: the position \( x_i \), the velocity \( v_i \) and the personal best position \( p_i \), where \( i \) denotes the number of particles and \( t \) denotes the iterations. The particle swarm has global best position \( g \), it is the best of personal best position. In each iteration, the position and the velocity are updated by the following equations.

\[
\begin{align*}
\dot{v}_i^{t+1} &= \omega v_i^t + \phi_1 R_1 (p_i^t - x_i^t) + \phi_2 R_2 (g^t - x_i^t) \quad (1) \\
\dot{x}_i^{t+1} &= x_i^t + v_i^{t+1} \quad (2)
\end{align*}
\]

where \( \omega \) denotes an inertia weight coefficient and \( \phi_1, \phi_2 \) are acceleration coefficients. \( R_1 \) and \( R_2 \) are randomly generated diagonal matrices. Each element of these matrices is a uniform random number in interval \([0, 1]\).

Rotation variance of PSO: The reason of the rotation variance of PSO is the search direction bias \([3]\). Figure 1 shows the histogram of the search direction and the trajectory of particles on 2-dimensional sphere function. The angle of the velocity vector means the search direction. Sphere function is isotropic. However, the search direction is biased in parallel to the coordinate axes.

In order to clarify the reason generating the bias of the search direction, we consider the simple velocity update rule that the reference position is one and without the inertia weight coefficient, as \( v_i^{t+1} = \phi R_i (p_i^t - x_i^t) \). Figure 2 shows the histogram of the 2-dimensional velocity vector when the angle of the reference position is fixed. Since the sign is not reversed, the distribution of the angle of the velocity vector is biased when the reference position vector is close to the coordinate axis.

3. Rotationally invariant PSOs

Several rotationally invariant PSOs have been proposed. In this section, we introduce the typical rotationally invariant PSOs.

Linear PSO (LPSO) is the most simple rotationally invariant PSO \([6][7]\). The velocity update rule of LPSO is described by the following equation.

\[
v_i^{t+1} = \omega v_i^t + \phi_1 r_1 (p_i^t - x_i^t) + \phi_2 r_2 (g^t - x_i^t), \quad (3)
\]
where \( r_1 \) and \( r_2 \) are uniform random numbers in the interval \([0, 1]\). In LPSO, the search direction is always pointed to the reference position. Since the random number is a scalar, the diversity of LPSO is poor.

**Rotation PSO (RPSO)** was proposed by Wilke et al. [6][7]. In RPSO, a rotation matrix is multiplied to the velocity vector.

\[
\begin{align*}
v_i^{t+1} &= \omega v_i^t + \phi_1 M_i (p_i^t - x_i^t) + \phi_2 M_2 (g^t - x_i^t) \\
M &= I + \frac{\alpha \pi}{180} (E - E^T)
\end{align*}
\]

where \( M \) is a random rotation matrix with the rotation angle \( \alpha \), and \( E \) is randomly generated matrix whose elements are uniform random numbers in the interval \([-0.5, 0.5]\). The dynamics of RPSO is closest to the dynamics of PSO. However, the calculation amount of the generating of the random rotation matrix is \( O(n^5) \). However, almost all elements of the rotation matrix are zero. By using the advantage of this fact, the calculation amount can be reduced to \( O(n^2) \). Based on this exact rotation matrix’s advantage, Bonyadi et al. proposed the method changes the rotation angle adaptively in the search process.

Clerc proposed **Standard PSO2011** (SPSO)[9] that realizes the rotation invariance by changing the shape of the search area. The velocity update rule of SPSO is described by the following equations.

\[
\begin{align*}
v_i^{t+1} &= \omega v_i^t + H(C_i^t, |C_i^t - x_i^t|) - x_i^t \\
C_i^t &= x_i^t + \frac{\phi_1 (p_i^t - x_i^t) + \phi_2 (g^t - x_i^t)}{3}
\end{align*}
\]

\( H(a, b) \) is a hypersphere function with the center \( a \) and the radius \( b \). Since the shape of the search area of SPSO is spherical, the biased search direction is not observed.

**Locally convergence rotationally invariant PSO (LcRiPSO)** was proposed by Bonyadi et al. [10]. LcRiPSO is the method combining the perturbed PSO [12] and LPSO. The random number of LcRiPSO is scalar as well as LPSO. However, since adding a normal random number to the reference position, the diversity of LcRiPSO is richer than LPSO.

\[
\begin{align*}
v_i^{t+1} &= \omega v_i^t + \phi_1 r_1 \left(N(p_i^t, (\sigma_1^t)^2 I) - x_i^t\right) + \phi_2 r_2 \left(N(g^t, (\sigma_2^t)^2 I) - x_i^t\right)
\end{align*}
\]

Bonyadi et al. proposed the method changes the variance \( \sigma_1^t, \sigma_2^t \) adaptively in the search process.

In NLPSO, the information of direction to the reference position is the sign only that is given by the sign function. Thus, the distribution of the angle of the velocity vector is not biased when the reference position vector is close to the coordinate axis.

\[
\begin{align*}
v_i^{t+1} &= \omega v_i^t + \phi_1 R_1 \left( ||p_i^t - x_i^t|| \right)\text{sing}(p_i^t - x_i^t) \\
&+ \phi_2 R_2 \left( ||g^t - x_i^t|| \right)\text{sing}(g^t - x_i^t)
\end{align*}
\]

Figure 4 shows the performance of PSO and rotationally invariant PSOs on 2-dimensional ellipse function [3]. The
The rotation variance of the PSOs on ellipse function.

Figure 4: The rotation variance of the PSOs on ellipse function.

Figure 5: The results of BBOB.

(a) Separable functions.  (b) All functions.

horizontally axis denotes the rotated angle of the coordinate system and the vertically axis denotes the median value in 25 trials. In each trial, the number of evaluations is 10000. From the result, the performance of PSO is degraded when the coordinate system is rotated. On the other hand, rotationally invariant PSOs are not dependent on the angle of the coordinate system. Since these experiments use only 2-dimensional functions, it is not possible to discuss the performance of these PSOs from the experimental results.

4. The performance of rotationally invariant PSOs

We investigate the performance of these PSOs by BBOB [13]. In order to evaluate the performance, we use the empirical cumulative distribution functions that are generated by COCO [14]. We set the recommended parameters. Figure 5 shows the results. In Fig 5, the horizontal axis denotes the log of the number of evaluation divided by the number of dimensions, and the vertical axis denotes the success rate in instances of each function.

Rotationally invariant PSO is not the method to improve performance but the method which resolved rotation variance. Thus, RPSO, MRPSO and NLPSO shows the similar performance as PSO. However, the results of all functions indicate that the performance of PSO is the best. Namely, resolving the search direction bias is the factor of deteriorating the performance. Because, in separable function, the biased search of PSO is advantageous. Thus, in Fig. 5a, the success rate of PSO is higher than rotationally invariant PSOs.

In a particularly high-conditioned and separable function, the biased search is advantageous. Figure 6 shows the performance of PSO and RPSO on the separable and non-separable convex function [5]. In these experiments, the maximum number of evaluations is $10^6$, and if the evaluation value reaches $10^{-4}$, until the maximum number of evaluations, this trial is regarded as a success. From the results, if the condition number is increased, the performance of RPSO is deteriorated. On the other hand, In the case of separable function, if the condition number is increased, the performance of PSO is hardly changed. However, in the case of non-separable function, if the condition number is increased, the performance of PSO is rapidly deteriorated.

5. A Novel PSO for high-conditioned and non-separable functions

In order to solve the high-conditioned and non-separable functions, we proposed new rotationally invariant PSO, it is described by the following equations [15].

$$v_{t+1} = \omega v_t + \phi_1 r_1 (p_{t} - x_t) + \phi_2 r_2 (g_t - x_t)$$  \hspace{1cm} (10)

$$p_t = p_t + c_d (p_{j1} - p_{j2})$$  \hspace{1cm} (11)

$$g_t = g_t + c_d (p_{j3} - p_{j4})$$  \hspace{1cm} (12)

$c_d$ is a constant number in the interval $[0, 1]$. $j_1, j_2, j_3$ and $j_4$ are random particle numbers, where $i \neq j_1 \neq j_2 \neq j_3 \neq j_4$. From the central limit theorem, the difference vector of the personal best follows the normal distribution. Thus, the proposed method is similar to LcRiPSO. However, the covariance matrix is different. In LcRiPSO, the covariance matrix of the perturbation becomes a diagonal matrix. On the other hand, the covariance matrix of the difference vector of the personal best is estimated as the inverse Hessian matrix of the objective function [15]. The estimation of the inverse Hessian matrix of the objective function is essential to solve the high-conditioned and non-separable functions [16]. Also, in order to improve the local search ability, we applied the selection mechanism [15].

To confirm the performance of proposed method, we carry out experiments. Table 1 shows the test functions. For each function, 25 trials are conducted. The parameter settings are refer to [15]. The time evolution of the best evaluation value in each trial is shown in Fig. 7.

From the results, the performance of the proposed method is better than the conventional PSO in the high-conditioned and non-separable functions.
Table 1: Test functions, where $y := Ax$ and $A$ is a rotation matrix.

<table>
<thead>
<tr>
<th>Name</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ellipsoid</td>
<td>$f_{\text{Ellipsoid}}(x) = \sum_{i=1}^{n} 10^6 (y_i^2 + y_i^2 + y_i^2 + y_i^2)^2 \left[ \sin^2 (50(y_i^2 + y_{i+1}^2)^0.1) + 1 \right]$</td>
</tr>
<tr>
<td>Rosenbrock</td>
<td>$f_{\text{Rosenbrock}}(x) = \sum_{i=1}^{n} 100(y_i^2 - y_{i+1}^2)^2 + (y_i - 1)^2$</td>
</tr>
<tr>
<td>Schaffer</td>
<td>$f_{\text{Schaffer}}(x) = \sum_{i=1}^{n} (y_i^2 + y_{i+1}^2)^{0.25} \left( \sin^2 (50(y_i^2 + y_{i+1}^2)^0.1) + 1 \right)$</td>
</tr>
</tbody>
</table>

6. Conclusions

In this article, we clarify that the factor of the rotation variance of PSO. We introduced typical rotationally invariant PSOs. Futhermore, in order to evaluate the performance of these PSOs, we carried out experiments using BBOB. From the result, the general performance of PSO is better than rotationally invariant PSOs. The reason is that rotationally invariant PSOs do not solve the separable/non-separable and high-conditioned functions. Also, in order to investigate the performance of proposed method, we carried out experiments. From the results, we clarified that our proposed method can solve the high-conditioned and non-separable functions.

Acknowledgment

This work was supported by JSPS KAKENHI Grant-in-Aid for Challenging Exploratory Research Number: 16K14271, Grant-in-Aid for Scientific Research(C) Number: 15K06077.

References


Figure 7: Time evolution of the best evaluation value for 10-dimensional test functions.
Nondeterministic Random Bits Extraction from Injected Chaotic Semiconductor Lasers

Xiao-Zhou Li1, Jun-Ping Zhuang1, Song-Sui Li1, and Sze-Chun Chan1,2,*

1Department of Electronic Engineering, City University of Hong Kong, Hong Kong, China
2State Key Laboratory of Millimeter Waves, City University of Hong Kong, Hong Kong, China
*Email: scchan@cityu.edu.hk

Abstract—Randomness extraction from an optically injected semiconductor laser is investigated. The generation of randomness from the chaotic intensity time series is examined by estimating the time-dependent exponents through state-space reconstruction. Chaotic dynamics enables fast divergence of neighboring states with positive exponents, while the possible effects of negative exponents have to be ruled out by using a sufficiently long sampling interval. This guarantees successful extraction of nondeterministic random bits at 200 Gbps from experimentally injected chaotic lasers.

1. Introduction
Fast physical random bit generation (RBG) is of great importance for a range of applications in computation and secure communication [1]. Despite the simplicity in using deterministic algorithms for generating pseudo-random bits, nondeterministic RBG processes have been extensively investigated using broadband photonic sources including quantum measurements of photons, optical noise, and optical chaos [1-12]. Amongst the different approaches for physical RBG, semiconductor laser chaos-based RBGs have attracted much attention as pioneered by Uchida et al. [4-12]. The chaotic dynamics typically produces fast temporal fluctuations that can be digitized for extraction into random bits at rates exceeding 1 Tbps [5]. Different schemes of chaos-based RBG have been reported using combinations of semiconductor lasers with optical feedback and optical injection [4, 9].

The randomness of such chaos-based RBG schemes can be examined by statistical evaluations using autocorrelation functions and power spectra [9, 13]. Practical tests from the National Institute of Standards and Technology (NIST) are also widely employed for verifying the randomness of the output bits [4]. These tests mainly focus on the statistical properties of the examined data, while the fundamental randomness for bits extraction is not guaranteed. Besides, it has been reported that pseudo-random bits from deterministic schemes can have sufficient quality for passing the NIST tests [14]. Therefore, randomness generation from chaotic semiconductor lasers needs to be investigated for nondeterministic RBG. Fundamentally, randomness is originated from the divergence of neighboring states through chaotic mixing. The divergence rate has been investigated in laser systems by estimating the largest Lyapunov exponents [13]. Randomness generation was also investigated by resetting the laser state repeatedly to the same initial state [15]. However, previously reported works on randomness evaluation for RBG are all based on semiconductor lasers with optical feedback, randomness extraction from an optically injected laser has not been much reported so far.

In this work, randomness extraction for chaos-based RBG is investigated using an optically injected semiconductor laser. Compared with optical feedback, chaos generated from optically injected lasers possesses the advantage of having no undesirable time-delay signatures that are commonly observed in time-delay systems [9, 10]. By numerically reconstructing the state space from an intensity time series, divergence of neighboring states is verified through the positive time-dependent exponents (TDEs). Negative exponents are also eliminated when the evolution time is sufficiently long. Furthermore, chaotic mixing is found to be essential for randomness generation, as compared with period-one (P1) dynamics. Successful RBG at 200 Gbps is demonstrated experimentally, as verified by the NIST tests. In particular, we examine the spread of the time-dependent exponent as a function of the initial distances between neighbors. Chaotic dynamics, as compared to P1 dynamics, is found to give much greater exponents with a significantly broader spread.

2. Schematic Setup

Fig. 1. Schematic of nondeterministic RBG using an optically injected semiconductor laser in chaos. ML, master laser; SL, slave laser; EDFA, erbium-doped fiber amplifier; CIR, circulator; PD, photodetector; ADC, analogue-to-digital converter.

The schematic of the setup for nondeterministic random bits extraction is shown in Fig. 1, where an optically injected semiconductor laser is utilized for chaos generation. The continuous-wave emission from a master laser ML is transmitted through an erbium-doped fiber amplifier EDFA and a circulator CIR for optically injecting a slave laser SL. Both ML and SL are single-mode semiconductor lasers. The injection parameters...
include the normalized injection strength \( \xi_i \) and the frequency detuning \( f_i \) of ML from SL. By adjusting the injection parameters \((\xi_i, f_i)\), SL can be driven into chaotic dynamics. After optical to electrical conversion by a photodetector (PD), it produces fast intensity fluctuations for digitization by analogue-to-digital converter (ADC) in an real-time oscilloscope, which is followed by digital processing for extraction into random bits.

For numerical simulations, the dynamics of a semiconductor laser can be described by the normalized complex intracavity optical field amplitude \( a(t) \) and the normalized charge carrier density \( \tilde{n}(t) \). With optical injection from ML, the dynamics of SL can be modeled by the following rate equations [16]:

\[
\begin{align*}
\frac{da}{dt} &= \frac{1}{2} \left[ \gamma_s a + \gamma_r |a|^2 - 1 \right] a \\
\frac{d\tilde{n}}{dt} &= -\left( \gamma_s + \gamma_r |a|^2 \right) \tilde{n} \\
&- \left( 1 - \frac{\gamma_r}{\gamma_c} |a|^2 \right) \left( |a|^2 - 1 \right),
\end{align*}
\]

where \( \gamma_s = 5.36 \times 10^{11} \text{s}^{-1} \) is the cavity decay rate, \( \gamma_r = 5.96 \times 10^{9} \text{s}^{-1} \) is the spontaneous carrier relaxation rate, \( \gamma_n = 7.53 \times 10^{9} \text{s}^{-1} \) is the differential carrier relaxation rate, \( \gamma_c = 1.91 \times 10^{10} \text{s}^{-1} \) is the nonlinear carrier relaxation rate, \( b = 3.2 \) is the linewidth enhancement factor, and \( J = 1.222 \) is the normalized bias current above threshold. The relaxation resonance frequency of the laser is \( f_r = 10.25 \text{ GHz} \). These parameters are extracted from a commercial semiconductor laser [16]. For simplicity, the spontaneous emission noise is not considered. Using second-order Runge-Kutta integration on Eqs. (1)–(2), an intensity time series \( I(t) \) is recorded with a sampling period of \( \tau_s = 2.38 \text{ ps} \). The injection parameters are chosen as \((\xi, f) \approx (0.05, 6.26 \text{ GHz}) \) for inducing chaotic dynamics in SL. The simulated chaos intensity is shown in Fig. 2(a).

Due to chaotic dynamics, the time series contains quick and irregular temporal fluctuations faster than 100 ps, which is comparable to the reciprocal of the relaxation resonance frequency of the laser.

3. Numerical State-space Reconstruction

To verify the divergence of neighboring states through chaotic mixing, the evolution of the trajectories needs to be examined. As only the emission intensity of the injected laser is usually measured in practical experiments, the reconstruction of the state space from the intensity time series is sought. For a normalized intensity time series \( I(t) \), a reconstructed state vector is given by \( \mathbf{x}(t) = [I(t), I(t + \tau_s), ..., I(t + (m_e - 1) \tau_s)] \), where \( m_e \) and \( \tau_s \) are the embedding dimension and embedding delay time, respectively. Since \( I(t) \) is usually recorded at a sampling period of \( \tau_s \), the \( i \)-th reconstructed state can be noted by \( \mathbf{x}_i = \mathbf{x}(i \tau_s) \), where \( i \) is the index of time. Suppose states \( \mathbf{x}_i \) and \( \mathbf{x}_j \) form a pair \((\mathbf{x}_i, \mathbf{x}_j)\) that describes two initial states. After an evolution time of \( \tau_s \) for some integer \( k \), the separation distance between the two states is given by \( d_{ij}(k) = ||\mathbf{x}_{i+k} - \mathbf{x}_{j+k}|| \), where \( || \cdot || \) denotes the Euclidean norm in the state space. So \( d_{ij}(0) \) is the initial distance between the two states. The evolution of the trajectories can be examined by the change of separation distance \( d_{ij}(k) \) over the evolution time \( k \tau_s \). Then the TDE is described by [17-19]:

\[
\Lambda_{ij}(k) = \frac{\ln d_{ij}(k)}{\ln d_{ij}(0)},
\]

The effect of divergence can be examined by the TDE when pairs of neighboring states are identified for \( d_{ij}(0) \) with sufficiently small values. Different pairs of neighbors give different initial separation distances \( d_{ij}(0) \) thus resulting in different TDEs \( \Lambda_{ij}(k) \) for a given evolution time \( k \tau_s \). It is of essence to scrutinize the effect of the choice of such initial distances on the statistics of the TDEs.

Figure 3(a) shows the so-called divergence plots for TDEs obtained from different pairs of neighboring states. An intensity time series of \( 10^7 \) data points is used for state-space reconstruction, where embedding parameters of \( m_e = 8 \) and \( \tau_s = 5 \tau_z \) are adopted. Similar reconstruction parameters have been utilized for optical injection chaos [18]. In Fig. 3(a-i), the evolution time is only 2.38 ps for \( k = 1 \). The time is too short for any pair of states to diverge, so most of the TDEs are concentrated at around 0. In Fig. 3(a-ii), the TDEs spread out when the evolution time increases to 0.05 ns for \( k = 21 \). Most of the neighboring states diverge due to chaotic dynamics, so their separation distances increase such that positive TDEs are observed. Negative TDEs are also identified, showing the existence of convergence between states. Interestingly, the TDEs spread more as the initial distance increases. This is because of associated increase of the number of initial states, which more completely probe the neighboring space. The evolution time further increases to 0.2 ns in Fig. 3(a-iii) for \( k = 84 \). Most of the neighbors keep diverging such that most TDEs continue to increase, while the number of negative TDEs also reduces. In Fig. 3(a-iv), as the evolution time is increased to 0.5 ns for \( k = 210 \), the TDEs generally keep increasing and nearly all TDEs become positive, as long as the evolution time is sufficiently long for the neighbors to diverge. Finally in Fig. 3(a-v), the evolution time is 1 ns for \( k = 420 \). The TDEs become more concentrated with positive values.
As the overall size of the chaotic attractor is not infinite, there is an upper bound for TDEs that decreases as the initial distance increases, as Figs. 3(a-iv) and 3(a-v) show. In fact, as the evolution time increases, the structure of the divergence plot becomes progressively invariant. The invariance is reached less quickly for states with small initial distances. The initially identified neighboring states, after the sufficiently long evolution time, are finally randomly located around the attractor. Therefore, any neighboring states are finally independent with each other, ensuring the fundamental randomness generation for RBG.

In order to examine the divergence of neighbors without chaotic mixing, Fig. 3(b) is shown by simulating the slave laser in P1 dynamics instead of chaotic dynamics. The P1 dynamics is obtained by adjusting the injection strength $\xi$ to 0.10. The emission intensity oscillates periodically at a microwave frequency of 16 GHz, as detailed in Fig. 2(b). The TDEs always stay at around 0, so nearby neighbors keep highly correlated even if the evolution time is long enough. Therefore, periodic oscillations without chaotic dynamics cannot support efficient divergence of neighboring states for randomness generation.

Summarizing Fig. 3, due to chaotic dynamics, TDEs spread out quickly and positively with the increase of time. Chaos is found to be essential for providing quick increase of TDEs for randomness generation.

Fig. 3. Time-dependent exponents for different pairs of neighboring states estimated from (a) chaotic and (b) P1 dynamics of an optically injected laser. The evolution time $k\tau_s$ increases for (i) $k = 1$, (ii) $k = 21$, (iii) $k = 84$, (iv) $k = 210$, and (v) $k = 420$.

4. Experimental RBG by Optical Injection

Based on the schematic setup detailed in Fig. 1, experiments are further conducted using a distributed-feedback semiconductor laser as the slave laser. In free-running, it emits at about 1550 nm with an optical power of 2.2 mW. The relaxation resonance frequency is $f_r = 7$ GHz. The maser laser ML is a tunable laser, and the detuning frequency is $f_i = 4$ GHz. The EDFA gain is controlled such that, with an injection power of 0.8 mW, chaotic dynamics is invoked in SL.

The emission light is converted to electrical signal through a PD, which is followed by a 13-GHz ADC with 8-bit resolution for digitization, where the sampling rate is set as 40 GHz. To ensure randomness, only 5 least significant bits are selected from each digitized sample. To reduce the bias introduced from imperfect detection, each sample is then compared with its 1-ns delayed replica through an exclusive-OR (XOR) operation. The output from the XOR is a stream of bits generated at an output bit rate of 200 Gbps. For comparison, by increasing the injection power to 1.6 mW, SL demonstrates P1 dynamics with its intensity oscillating at 9 GHz.

Then RBGs by chaotic dynamics and P1 dynamics are compared experimentally through taking the NIST tests. A total of 1000 sequences, each of size 1 Mbit, are collected for testing. At significance level $\alpha = 0.01$, the success proportion has to be in the range of $0.99 \pm 0.0094392$ for passing a test. The composite $P$-value should be larger than 0.0001 to ensure uniformity. The testing results for chaos and P1 are respectively summarized in Tables 1 and 2, where the worst case is shown for tests producing multiple $P$-values and proportions.

Random bits generated by chaotic dynamics can successfully pass all the 15 NIST tests, while most of the tests are failed using P1 dynamics. This again verifies the randomness generation requires using chaotic dynamics of an optically injected semiconductor laser.

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>$P$-value</th>
<th>Proportion</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0.007862</td>
<td>0.9810</td>
<td>Success</td>
</tr>
<tr>
<td>Block-frequency</td>
<td>0.206629</td>
<td>0.9920</td>
<td>Success</td>
</tr>
<tr>
<td>Cumulative-sums</td>
<td>0.007975</td>
<td>0.9820</td>
<td>Success</td>
</tr>
<tr>
<td>Runs</td>
<td>0.725829</td>
<td>0.9850</td>
<td>Success</td>
</tr>
<tr>
<td>Longest-run</td>
<td>0.415422</td>
<td>0.9860</td>
<td>Success</td>
</tr>
<tr>
<td>Rank</td>
<td>0.773405</td>
<td>0.9870</td>
<td>Success</td>
</tr>
<tr>
<td>FFT</td>
<td>0.204439</td>
<td>0.9900</td>
<td>Success</td>
</tr>
<tr>
<td>Nonoverlapping-templates</td>
<td>0.340858</td>
<td>0.9820</td>
<td>Success</td>
</tr>
<tr>
<td>Overlapping-templates</td>
<td>0.695200</td>
<td>0.9900</td>
<td>Success</td>
</tr>
<tr>
<td>Universal</td>
<td>0.927677</td>
<td>0.9920</td>
<td>Success</td>
</tr>
<tr>
<td>Approximate-entropy</td>
<td>0.769527</td>
<td>0.9920</td>
<td>Success</td>
</tr>
<tr>
<td>Random-excursions</td>
<td>0.478196</td>
<td>0.9834</td>
<td>Success</td>
</tr>
<tr>
<td>Random-excursions-variant</td>
<td>0.158133</td>
<td>0.9834</td>
<td>Success</td>
</tr>
<tr>
<td>Serial</td>
<td>0.655854</td>
<td>0.9930</td>
<td>Success</td>
</tr>
<tr>
<td>Linear-complexity</td>
<td>0.637119</td>
<td>0.9820</td>
<td>Success</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>15</strong></td>
</tr>
</tbody>
</table>

Table 1. NIST test results for RBG using chaos from an optically injected semiconductor laser.
Table 2. NIST tests results for RBG using P1 from an optically injected semiconductor laser

<table>
<thead>
<tr>
<th>Statistical test</th>
<th>P-value</th>
<th>Proportion</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>0.000000</td>
<td>0.0450</td>
<td>Fail</td>
</tr>
<tr>
<td>Block-frequency</td>
<td>0.000000</td>
<td>0.0000</td>
<td>Fail</td>
</tr>
<tr>
<td>Cumulative-sums</td>
<td>0.000000</td>
<td>0.0380</td>
<td>Fail</td>
</tr>
<tr>
<td>Runs</td>
<td>0.000000</td>
<td>0.0000</td>
<td>Fail</td>
</tr>
<tr>
<td>Longest-run</td>
<td>0.000000</td>
<td>0.8940</td>
<td>Fail</td>
</tr>
<tr>
<td>Rank</td>
<td>0.842937</td>
<td>0.9860</td>
<td>Success</td>
</tr>
<tr>
<td>FFT</td>
<td>0.000000</td>
<td>0.6180</td>
<td>Fail</td>
</tr>
<tr>
<td>Nonoverlapping-templates</td>
<td>0.000000</td>
<td>0.6150</td>
<td>Fail</td>
</tr>
<tr>
<td>Overlapping-templates</td>
<td>0.000000</td>
<td>0.6530</td>
<td>Fail</td>
</tr>
<tr>
<td>Universal</td>
<td>0.000000</td>
<td>0.0710</td>
<td>Fail</td>
</tr>
<tr>
<td>Approximate-entropy</td>
<td>0.000000</td>
<td>0.0000</td>
<td>Fail</td>
</tr>
<tr>
<td>Random-excursions</td>
<td>0.006582</td>
<td>0.9846</td>
<td>Success</td>
</tr>
<tr>
<td>Random-excursions-variant</td>
<td>0.000000</td>
<td>0.0000</td>
<td>Fail</td>
</tr>
<tr>
<td>Serial</td>
<td>0.000000</td>
<td>0.0000</td>
<td>Fail</td>
</tr>
<tr>
<td>Linear-complexity</td>
<td>0.222480</td>
<td>0.9910</td>
<td>Success</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

5. Conclusion

In conclusion, nondeterministic RBG is investigated using a chaotic optically injected semiconductor laser. By estimating the TDEs from a reconstructed state-space, the divergence of neighboring states by chaotic dynamics is illustrated through a divergence plots (Fig. 3). Chaotic dynamics is also found to be essential for randomness generation, as compared with PI dynamics, in the experiments at 200 Gbps.

Acknowledgments

The work described in this paper was fully supported by a grant from the National Natural Science Foundation of China (NSFC) (Grant 61308002).

References

Fast physical random bit generation by chaotic lasers with delayed feedback using extremely short external cavities

Shin Suzuki¹, Satoshi Sunada², Susumu Shinohara³, Takehiro Fukushima⁴, and Takahisa Harayama¹

¹Department of Applied Physics, School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
²Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan
³NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
⁴Department of Information and Communication Engineering, Okayama Prefectural University, 111 Kuboki, Soja, Okayama 719-1197, Japan
Email: sin-yougaku@akane.waseda.jp

Physical (true) random bits play an essential role in secure communications and data encryptions. As a device for physical random bit generation, chaotic lasers with delayed optical feedback have attracted considerable attention, because they make possible fast (gigabits/second) physical random bit generation. Such a chaotic laser has recently been realized by a photonic integrated circuit on a chip [1]. The laser chaos chip consists of a photodiode, a semiconductor distributed feedback laser, semiconductor optical amplifiers, and a passive waveguide for delayed optical feedback. The length of the passive waveguide is 1 cm, which is almost equal to the whole system size, since the sizes of all the other optical components are much less than 1 mm. Although a shorter passive waveguide is desirable for device size reduction, it is not clear how short we can make it, while maintaining the capability of generating good quality of random bit at the fast rate of gigabits/second.

In this work, we numerically studied random bit generation in cases of short passive waveguides by using the Lang-Kobayashi equations. The generated random bit sequences were evaluated by the statistical tests of randomness provided by NIST [2]. We found that by controlling the injection current and the amount of the optical feedback, we can obtain highly chaotic oscillations, even when the lengths of the passive waveguides are much shorter than 1 cm. Using these chaotic oscillations, we were able to generate random bits at the rate of gigabits/second that pass all of the NIST tests.

Optimal design of two-dimensional external cavities for delayed optical feedback

Shinji Orihara¹,*, Koei Koyama¹,#, Susumu Shinohara², Satoshi Sunada³, Takehiro Fukushima⁴, and Takahisa Harayama¹

¹Department of Applied Physics, School of Advanced Science and Engineering, Waseda University, 
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan 
²NTT Communication Science Laboratories, NTT Corporation, 
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan 
³Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University, 
Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan 
⁴Department of Information and Communication Engineering, Okayama Prefectural University, 
111 Kuboki, Soja, Okayama 719-1197, Japan

Email: *kh5462@asagi.waseda.jp, #poohtaro-koei@toki.waseda.jp

Recently, chaotic semiconductor lasers with delayed optical feedback have attracted renewed interest, because of their usefulness as an entropy source for physical random number generation [1]. It has been reported that, for obtaining strongly chaotic outputs suitable for random number generation, a sufficiently long external cavity (e.g. 2 mm) for delayed optical feedback is necessary. This requirement is an obstacle to the miniaturization of the device. As an idea to resolve this problem, folding an optical path in a two-dimensional (2D) external cavity has been proposed [2], where the cavity consists of a deformed microdisk and two linear waveguides. This cavity makes it possible to form a path much longer than the cavity diameter [see Fig.1 for the cavity shape]. For example, a 2.8-mm path was formed for a device with the diameter of 0.3 mm [2].

In this approach, the cavity shape needs to be carefully designed to suppress light diffusion due to the multiple reflections. In this work, we are interested in an optimal cavity shape that maximizes the feedback strength. We simulated the light propagation in the 2D external cavity by the flux of rays, and determined an optimal cavity shape where the maximal feedback strength is expected. By performing wave simulations based on the FDTD method, we confirmed the improvement of the feedback strength up to 8.9% as compared to the previously reported cavity shape.

Fig.1: 2D external cavity with a folded long path.

Intermittent dynamics in lasers: Distribution mapping and random number generation

Andreas Karsaklian Dal Bosco\textsuperscript{1}, Naoki Sato\textsuperscript{1}, Shoma Ohara\textsuperscript{1}, Yuta Terashima\textsuperscript{1}, Atsushi Uchida\textsuperscript{1}, Takahisa Harayama\textsuperscript{2,3} and Kazuyuki Yoshimura\textsuperscript{2}

\textsuperscript{1} Department of Information and Computer Sciences, Saitama University
255 Shimo-Okubo Sakura-ku, Saitama City, Saitama 338-8570, Japan
\textsuperscript{2} NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato, Wakamiya, Atsugi-Shi, Kanagawa 243-0198, Japan
\textsuperscript{3} Department of Applied Physics, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
Emails: andreaskdb@gmail.com, auchida@mail.saitama-u.ac.jp

Abstract—We study the properties of intermittent dynamics occurring in semiconductor lasers embedded in five photonic integrated circuits subjected to optical feedback with different delay times. Intermittency is a transition between two different dynamics induced by a change in the value of a control parameter (feedback strength, injection current). Intermittent waveforms exhibit the features of both dynamics, distributed in a succession of laminar regions and bursts. Typical intermittent dynamics puts into play steady states or low-amplitude periodic dynamics in the laminar regions while bursts are made of chaos or high-amplitude oscillations. We show how the distribution of intermittency is affected as the injection current and feedback strength in the lasers are varied. This is done by mapping all the dynamics observed in two-dimensional bifurcation diagrams corresponding to the five values of feedback delay times (scaling from 33 to 265 ps). We also propose a method based on the irregularity ruling the succession of the laminar regions and bursts specific to intermittency to generate random numbers.

1. Introduction

The recent development of photonic integrated circuits (PICs) has led to major technological advances in optics, allowing in particular to reduce the spatial extension of optical devices considerably [3]. Monolithic PICs have been designed to yield optical chaos to be used for applications in telecommunications [4], fast random bit generation [5, 6, 7, 8] and all-optical self-pulsation generation [9, 10]. The main advantage that PICs offer is the possibility to implement different optical functions such as light emission, amplification and detection in a single device having a minimal spatial extension (sub-centimeter scale). This property opens the way to investigations of the dynamical diversity observed in semiconductor lasers [11].

We report different organizations of the dynamics yielded in semiconductor lasers embedded in PICs subjected to different optical feedback delay times (ranging from 33 to 265 ps). We focus on the variation of the distribution of the dynamics the laser injection current, the feedback strength and the feedback delay time are varied. We first present the how changes in the external cavity length can induce different distributions of the dynamics in given ranges of injection current and feedback strength. Then we focus on a particular dynamics, termed intermittency, and explain how its existence is conditioned by the feedback delay time. Intermittency is a dynamical behavior in which the time trace is organized into laminar regions and bursts, successively following each other in time. We explain the mechanism of this intermittent dynamics and its role in the bifurcation diagrams. Finally, we propose a method to generate random numbers based on the intermittent dynamics, taking advantage of the temporal randomness ruling the irregular successions of laminar regions and bursts. This method differs from the traditional method based on the irregularities of the amplitude of chaotic waveforms [2], with which we bring a comparison of the performances.

2. Experimental setup and observations

Each PIC used in our research consists of a semiconductor laser bounded by a photodiode and an external cavity composed of two independent semiconductor optical amplifiers and a passive waveguide ended by a reflector. The structure of the PICs is presented in Fig. 1. The external cavity lengths range from 1.3 to 10.3 mm, giving external cavity frequencies between approximately 4 and 30 GHz.

In order to understand the distribution of the various dynamics observed in the PICs, an experimental two-dimensional bifurcation diagram obtained by varying the feedback strength $J_{SOA}$ and injection current $J/J_{th}$ for a PIC having a 3.3-mm external cavity length is presented in Fig. 2. The colors correspond to the different behaviors observed when changing these two parameters. The laser exhibits a large variety of dynamics with regions of steady states, periodic, quasi-periodic dynamics, chaos and low-frequency fluctuations, as can usually be seen in regular
3. Random number generation based on intermittent dynamics

The irregular temporal distribution of the laminar regions and bursts in the temporal waveforms showing intermittent dynamics motivates to use this kind of dynamics to generate random numbers. By contrast to traditional methods consisting in sampling the amplitude of a chaotic output waveform, we propose in the present case to use the temporal randomness ruling the succession of consecutive bursts as a source for random numbers. The principle we implement consists in a first time in counting all the times of the laminar regions (times between consecutive bursts). Then all the obtained values are converted into an equivalent number of sampled points, by taking into account the sampling rate of the oscilloscope used for the temporal acquisitions. The following step is a $2n$ modulo operation applied to these numbers of sampled points. The lengths of the laminar regions are defined as the remainders obtained in this operation, in which $n$ is an integer between 1 and 16. As a final step, the values of the laminar times are changed into binary sequences of $n$ bits. The method used for sampling the laminar times is illustrated in Fig. 4.

In order to study the performances of the random number generation process by using this method, we carry out a comparison between this direct method and two enhanced methods, using respectively the XOR and bit order reverse methods. We demonstrate that the random number genera-
Figure 4: Method for generating random number from a direct conversion of the times of the laminar lengths of the temporal waveform.

4. Conclusion

We presented an experimental analysis of the distribution of the different dynamics that can be yielded in photonic integrated circuits with optical feedbacks from external cavities of a few millimeters. Under the effect of the injection current and the feedback strength, the two-dimensional mapping of the lasers show different properties. In particular, the presence of intermittent dynamics accompanying the stabilization and destabilization of chaotic dynamics has been evidenced for different feedback delay times. We also proposed a method for generating random numbers from optical intermittent dynamics based on the temporal randomness governing the cadency at which bursts rise in the temporal waveform of the laser emitted power. By comparing several post-process methods with different sets of parameters, we demonstrated that random number generation was possible with this method.

References


Fast physical random bit generation using a photonic integrated circuit

Yuta Terashima¹, Kazusa Ugajin¹, Atsushi Uchida¹, Takahisa Harayama²,³ and Kazuyuki Yoshimura²,⁴

¹ Department of Information and Computer Sciences, Saitama University
255 Shimo-Okubo Sakura-ku, Saitama City, Saitama 338-8570, Japan
² NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato, Wakamiya, Atsugi-Shi, Kanagawa 243-0198, Japan
³ Department of Applied Physics, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
⁴ Department of Information and Electronics Graduate school of Engineering, Tottori University
4-101 Koyama-Minami, Tottori 680-8552, Japan

Emails: s15mm318@mail.saitama-u.ac.jp, auchida@mail.saitama-u.ac.jp

Abstract—We propose random bit generation by using laser chaos generated from a photonic integrated circuit. We obtain smooth and symmetrical distribution of the amplitude of a chaotic temporal waveform by using a differential method. We generate random bit sequence by applying the differential method and XOR-operation, and the generated bits pass all of the NIST Special Publication 800-22 tests and TestU01 tests.

1. Introduction

Random number generation based on physical phenomena is useful for information security and cryptographic communication. Recently, random number generation using chaotic semiconductor lasers has been intensively investigated and high generation rates over Gigabit per second (Gb/s) have been reported [1–6]. In these systems, chaotic generators are usually composed of several devices such as a laser, a photodetector, and an external mirror. Therefore, photonic integrated circuits (PICs) have been proposed to reduce the size of physical random number generators [2, 3].

The statistical histogram of chaotic signals typically shows asymmetric distribution for laser-chaos-based random bit generation. The quality of the randomness of the random numbers generated from the chaotic signal is degraded due to the asymmetric distribution. To solve this problem, a differential method has been proposed to produce symmetric statistical distribution of the chaotic signals [4, 5].

To evaluate the randomness of the random bit sequence, NIST Special Publication 800-22 (NIST SP 800-22) have been commonly used. The length of the random number by the NIST evaluation requires only 1 Gigabits (10⁹ bits). The statistical tests that treat with larger amounts of random numbers are necessary, and one of these statistical tests is known as TestU01 [9]. The maximum random bit length required for TestU01 is about 410 Gigabits.

In this study, we experimentally generate random bit sequences from chaotic temporal waveforms obtained from a PIC by using a differential method and exclusive-or (XOR) operation. We evaluate large amount of random bit sequences by using NIST SP 800-22 and TestU01.

2. Experimental setup

Figure 1 shows the schematics of the PIC used for random bit generation. The PIC consists of a semiconductor laser, two semiconductor optical amplifiers, a 10-mm-long waveguide, an external mirror, and a photodetector. A chaotic laser output is generated by the optical feedback from the mirror. The feedback strength can be tuned by varying the injection currents of the optical amplifiers. The chaotic optical signal of the laser is converted into an electrical signal by the photodetector.

Figure 2 shows the experimental setup for random bit generation. The temporal waveform of the laser output from the PIC is divided into the alternating current (AC) and the direct current (DC) components by a bias tee. The AC component is amplified by using an electrical amplifier, and sampled by an analog-to-digital converter with 8-bit resolution.

3. Experimental results of chaotic temporal waveform

Figure 3 shows a typical example of the chaotic temporal waveform and the radio-frequency (RF) spectrum generated from the PIC. An irregular temporal waveform is...
observed in Fig. 3(a). A broadband spectrum is obtained in Fig. 3(b).

Figure 4 shows the probability distributions of the original temporal waveform and that generated from the difference between the original and time-delayed temporal waveforms by using the differential method. The histogram is obtained by sampling the amplitude of the chaotic temporal waveform with the vertical resolution of 256 points. The histogram shows irregularities due to the quantization error of the AD converter as shown in Fig. 4(a). Such a quantization error can cause artificial randomness of the generated random bits. To eliminate these quantization errors, we apply the differential method to the chaotic signal [4, 5]. The histogram after applying the differential method shows a smooth and normal distribution as shown in Fig. 4(b).

4. Random bit generation method

We generated a random bit sequence from the chaotic temporal waveforms by using the post-processing, consisting of the differential method, the bit-order reversal, and XOR operation. First, we apply a differential method to the chaotic signal to prevent the undesired asymmetric distribution of the histogram of the laser intensities, as shown in Fig. 5. Next, we generate three time-delayed 8-bit signals (called D1, D2, and D3) from the original signal (called D0). Then, we generate two differential signals (called DS1 and DS2) by calculating the differences between D0 and D1, and between D2 and D3 respectively. It is expected that the statistical histograms of DS1 and DS2 show symmetric distributions. Next, the bit order of DS2 is reversed, i.e., the most significant bit (MSB) changes to the least significant bit (LSB), the second MSB changes to the second LSB and so on [6] as shown in Fig. 6. Bitwise exclusive-OR (XOR) operation is then carried out between the bit-order reversed DS2 and the original DS1. Finally, some of the LSBs are extracted from the 8-bit signal, and they are used as a random bit sequence.

5. Evaluation of generated random bits

To evaluate the randomness of a long random bit sequence, we calculate the statistical bias $b$ of the occurrence of bit ‘1’. The bias $b$ is defined as follows.

$$b = |p(1) - 0.5|$$ (1)
where \( p(1) \) is the probability of the occurrence of ‘1’. Smaller bias indicates higher randomness of generated random bit sequences. For the finite length \( N \) of the random bit sequence, the statistical bias \( b \) needs be less than the three-standard-deviations, defined as \( 3\sigma = 1.5N^{-0.5} \) [7]. We calculated the statistical bias \( b \) as a function of the length of the generated bit sequence \( N \), and confirm that the bias is less than the \( 3\sigma \) line for all \( N \).

The generated random bits are evaluated using National Institute of Standards and Technology Special Publication 800-22 (NIST SP 800-22) [8]. The NIST tests are performed for 1000 sequences of 1 Mbit length. Typical results of the NIST tests are shown in Table 1. The random bit sequence generated from the post-processing with 8 LSBs passes all of the NIST tests.

Figure 7 shows the number of the passed NIST tests when the number of extracted LSBs is changed. We carried out the NIST tests with 1-Gbit sequences for five times, and the median of the five test results is plotted with error bars of the maximum and minimum values, as shown in Fig. 7. We succeeded in passing all of NIST tests in the cases from 1 to 8 LSBs except LSB 3.

Table 1: Result of NIST SP 800-22 for random bit sequences generated from the post-processing with 8 LSBs. Significance level is set to \( \alpha = 0.01 \). To pass the tests, the \( P \)-value of the uniformity of \( p \)-values should be larger than 0.0001, and the proportion of sequences satisfying \( P \)-value > \( \alpha \) for 1000 samples of 1 Mbit data should be in the range of 0.99 ± 0.0094392 [8]. For tests which produce multiple \( P \)-values and proportions, the worst case is shown.

<table>
<thead>
<tr>
<th>Number of extracted LSBs</th>
<th>Number of passed tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15/15</td>
</tr>
<tr>
<td>2</td>
<td>15/15</td>
</tr>
<tr>
<td>3</td>
<td>15/15</td>
</tr>
<tr>
<td>4</td>
<td>15/15</td>
</tr>
<tr>
<td>5</td>
<td>15/15</td>
</tr>
<tr>
<td>6</td>
<td>15/15</td>
</tr>
<tr>
<td>7</td>
<td>15/15</td>
</tr>
<tr>
<td>8</td>
<td>15/15</td>
</tr>
</tbody>
</table>

Figure 7: Number of passed NIST tests as a function of the extracted LSBs for the post-processing for the random bit generation. “15” on the vertical axis indicates that all the NIST tests are passed. Five 1-Gbit sequences of random bits are used for each NIST test and the median of the five test results is plotted with error bars of the maximum and minimum values.

6. Conclusions

We experimentally demonstrated fast physical random bit generation by using laser chaos generated from a photonic integrated circuit. We generated random bit sequence by applying the differential method, bit-order reversal, and XOR operation. The randomness of the generated random bit sequences is verified by using NIST Special Publication 800-22 and TestU01.

Acknowledgments

We acknowledge support from Grants-in-Aid for Scientific Research from Japan Society for the Promotion of Science (JSPS KAKENHI Great Number JP24686010), and Management Expenses Grants from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

We also used TestU01 to evaluate the randomness of a longer random bit sequence. TestU01 consists of five packages: Rabbit, Alphabit, SmallCrush, Crush, and BigCrush. The BigCrush test requires the longest random bit length of \( \sim 410 \) Gbits (\( 4.1 \times 10^{11} \) bits). We carried out the TestU01 tests while the number of the extracted LSBs is changed. It found that all of the Rabbit, Alphabit, and SmallCrush tests are passed from 1 to 8 LSBs, and all of the Crush and BigCrush test are passed from 5 to 8 LSBs.

References


Stability Analysis of Periodic Orbits in Dynamic Binary Neural Networks

Shunsuke Aoki†, Ryuji Sato †, Kazuma Makita †, and Toshimichi Saito †

†Department of Electronics and Electrical Engineering, Hosei University
3–7–2 Kajino-cho, Koganei, Tokyo 184-8584, Japan
Emails: {shunsuke.aoki.7x, kazuma.makita.4w, ryuji.sato.5b}@stu.hosei.ac.jp; tsaito@hosei.ac.jp

Abstract—This paper studies various periodic orbits and their stability in dynamic binary neural networks with signum activation function. The network can generate various binary periodic orbits and the dynamics is integrated into a digital return map defined on a set of points. Performing basic numerical experiments, it is shown that the network can generate various periodic orbits and ternary \{-1, 0, +1\} connection parameters can reinforce stability of the periodic orbits.

1. Introduction

The dynamic binary neural network (DBNN) is constructed by applying a delayed feedback to a feedforward binary neural network with signum activation [1]-[5]. Depending on connection parameters, the DBNN can generate a variety of binary periodic orbits (BPOs). The DBNN is included in digital dynamical systems such as cellular automata [8] and digital spiking neurons [10]. These digital dynamical systems are applicable to logical/sequential circuits, image processing systems, and UWB communication systems. Analysis of DBNN is important from both fundamental and application viewpoints. However, analysis of DBNN is hard because the DBNN has a large number of parameters and can generate a large variety of periodic/transient phenomena.

This paper analyzes dynamics of a simple class of DBNN. First, we introduce that the dynamics of the DBNN is integrated into a digital return map (Dmap) from a set of lattice points to itself. The Dmap can be regarded as a digital version of analog return map represented by the logistic map [11]. We then analyze three types of 6-dimensional DBNNs: the connection parameters $w_{ij}$ are integer, they are binary $w_{ij} \in \{-1, 0, +1\}$, and they are ternary $w_{ij} \in \{-1, 0, +1\}$. Such 6-dimensional DBNNs are applicable to control signal of basic dc/ac and ac/dc power converters [3] [5].

Performing numerical experiments it is shown that (1) the binary connection parameters are able to realize the almost same stability of BPO as the integer connection parameters, (2) the ternary connection parameters are able to reinforce stability of the BPO, and (3) the binary and ternary connection parameters can suppress spurious memories.

2. Dynamic Binary Neural Networks

The DBNN is constructed by applying a delayed feedback to a feed-forward network with the signum activation function. The dynamics is described by

$$x_{i}^{t+1} = \text{sgn} \left( \sum_{j=1}^{N} w_{ij} x_{j}^{t} - T_{i} \right)$$

where $x^t$ is a binary state vector at discrete time $t$ and $x_{i}^{t} \in \{-1, 0, +1\} \equiv B$ is the $i$-th element. The connection parameters $w_{ij}$ and the threshold parameters are integer.

In order to visualize the dynamics DBNN, we introduce the Dmap. The domain of the DBNN is a set of

![Figure 1: (a) DBNN. $w_{ij} = 0$ means no connection. The threshold parameters $T_{i}$ are shown in the circles. (b) Dmap. DBNN and Dmap. Red orbit denotes TBPO. Green orbit denotes a BPO. $\gamma = 10/16$](image-url)
binary vectors $B^N$ that is equivalent to a set of points
$L_D = \{C_1, \ldots, C_{2N}\}$. Since the dynamics of the
DBNN can be integrated into the digital return map (Dmap)
from $L_D$ to itself:

$$x^{r+1} = F_D(x^r), \quad x^r \equiv (x_i^r, \ldots, x_{2N}^r) \in B^N$$ (2)

Figure 1 illustrates the Dmap for $N = 4$ where bi-

cinary code is used to express $L_4 = \{C_1, \ldots, C_{16}\}$: $C_1 \equiv
\cdots C_{16} \equiv (\pm 1, \pm 1, \pm 1, \pm 1)$. Since the

number of lattice points is $2^N$, direct memory of all the

inputs/outputs becomes hard/impossible as $N$ increases.

However, in the DBNN, the number of parameters is poly-
nomial $N^2 + N$.

Since the number of the lattice points is finite, the steady

states are BPOs defined as the following.

A point $\theta_{p} \in L_D$ is said to be a periodic point (PEP)

with period $p$ if $F^p(\theta_p) = \theta_p$ and $F^k(\theta_p) \neq \theta_p$ for $1 \leq k < p

where $F^k$ is the p-fold composition of $F$. Especially, a point

with period 1 is said to be a fixed point. A sequence of the

PEPs, $(F(\theta_p), \ldots, F^p(\theta_p))$, is said to be a binary periodic

orbit (BPO).

In Fig. 1 the Dmap has one BPO with period 3 and BPO

with period 2. Depending on the initial condition, the

DBNN exhibits either BPO.

3. Teacher signal and stability

We consider storage of one BPO into the DBNN. The

teacher signal binary periodic orbit (TBPO) with period $T

is described by

$$z^1, z^2, \ldots, z^T, \quad \quad \quad \quad \quad z^i = (z_i^1, \ldots, z_i^N) \in B^N$$

$$z^i = z^j \quad \text{for} \quad |i - j| = T, \quad z^i \neq z^j \quad \text{for} \quad |i - j| \neq T$$ (3)

For simplicity, the period $T$ is assumed to be order of $N$. In

order to determine connection parameters $w_{ij}$, we use three

methods A parameter condition for storage of TBPO.

Integer connection: $w_{ij} = c_{ij} = \sum_{z=1}^{T} z_i^{r+1} z_j^{r}$ (4)

Binary connection: $w_{ij} = b_{ij} = \begin{cases} +1 & \text{for } c_{ij} \geq 0 \\ -1 & \text{for } c_{ij} < 0 \end{cases}$ (5)

Ternary connection: $w_{ij} = d_{ij} \in \{1, 0, +1\}$

d$ij$ is given by zero-insertion algorithm in [6].

Ref. [6] gives a sufficient condition of parameters for stor-

age of TBPO. Referring to the condition, as connection pa-
neters $w_{ij}$ are given, the threshold parameters $T_i$ can be
determined theoretically.

Here we define stability of TBPO. A TBPO is said to

be stable if at least one initial point (except for the TBPO)
falls into the TBPO. A TBPO is said to be globally stable

if all initial points into the TBPO. In order to characterize

the global stability, we introduce a simple feature quantity

$$\gamma = \# \text{Initial points falling into the TBPO} \frac{2^N}{T}$$ (7)

where $T/2^N \leq \gamma \leq 1$. If $\gamma = 1$ then the TBPO is globally

stable. In Fig. 1, 10 initial points fall into the TBPO with

period 3 and $\gamma = 10/16$.

4. Numerical experiment

This paper considers two examples of TBPOs with pe-

riod 6. For $N = 6$, Table 1 shows the first TBPO corre-

sponding to a control signal of a basic AC/DC converter.

Applying the three kinds of connection parameters in Eqs.

(4) to (6), the first TBPO can be stored into the DBNN.

Tables 3, 4, and 5 show integer, binary, and ternary connection

parameters, respectively. Figures. 2, 3, and 4 show cor-

responding three Dmaps. The integer, binary, and ternary

connection parameters give $\gamma = 17/64$, $\gamma = 12/64$, and

$\gamma = 1$, respectively.

Table 2 shows the second TBPO corresponding to a con-

rol signal of a basic AC/DC converter. Applying the three

kinds of connection parameters in Eqs. (4) to (6), the sec-

ond TBPO can be stored into the DBNN. Tables 6, 7, and

8 show integer, binary, and ternary connection parameters,

respectively. Figures. 5, 6, and 7 show corresponding three

Dmaps. The integer, binary, and ternary connection parameters

give $\gamma = 42/64$, $\gamma = 42/64$, and $\gamma = 1$, respectively.

In these results, we can see the following.

- The binary connection can realize the almost same

  global stability as the integer connection.

- The ternary connection with zero elements can realize

  global stability. The global stability is impossible in

  integer and binary connections.

- The binary and ternary connections can suppress seri-

  ous memories in the case of integer connection.
5. Conclusions

A class of 6-dimensional DBNNs has been studied in this paper. In order to visualize the DBNN dynamics, the Dmap is introduced. In order to consider the stability of TBPO, global stability is defined and a simple feature quantity $\gamma$ is introduced.

In basic numerical experiments, we have shown that the ternary connection can realize global stability and the binary/ternary connection can suppress spurious memories.

Future problems include analysis of 2DBNN with sparse connections, analysis of deep DBNNs, and engineering applications.

References


Figure 2: Dmap for Table 3. Red orbit: TBPO. Green orbit: Spurious BPO. $\gamma = \frac{17}{64}$

Figure 3: Dmap for Table 4. Red orbit: TBPO. Green orbit: Spurious BPO. $\gamma = \frac{12}{64}$

Figure 4: Dmap for Table 5. Red orbit: TBPO. $\gamma = 1$

Figure 5: Dmap for Table 6. Red orbit: TBPO. Green orbit: Spurious BPO. $\gamma = \frac{42}{64}$

Figure 6: Dmap for Table 7. Red orbit: TBPO. Green orbit: Spurious BPO. $\gamma = \frac{42}{64}$

Figure 7: Dmap for Table 8. Red orbit: TBPO. $\gamma = 1$
Multi-compartment Neuron Model based on Asynchronous Bifurcation Processor

Taiki Naka† and Hiroyuki Torikai‡

†Department of Computer Science, Kyoto Sangyo University
Motoyama, Kamigamo, Kita-Ku, Kyoto 603-8555, Japan
g1344900@cse.kyoto-su.ac.jp, torikai@cse.kyoto-su.ac.jp

Abstract—In this paper, a multi-compartment neuron model based on the concept of an asynchronous bifurcation processor is studied. It is shown that the model can reproduce typical propagation phenomena of membrane potentials between somas and dendrites of neurons such as forward propagations of action potentials from dendrites and resulting backward propagations of action potentials from a soma. It is also shown that the model can reproduce formations of typical potential gradients in dendrites of neurons.

1. Introduction

A neuron typically consists of dendrites (input cables), a soma (cell body), and axons (output cables), where the dendrite sometimes has complicated physical structure such as the Purkinje cell. A wide variety of dendritic phenomena have been observed [1]-[4], where it has been suggested that such dendritic phenomena play certain roles in neural information processing as well as spike-timing dependent plasticity learning. One of the major modeling methods of the “soma plus dendrite” is a multi-compartment soma-dendrite modeling method, i.e., to discretize the dendrite into a set of small compartments and to model the “soma plus dendrite” by a coupled system of the compartments as shown in Fig. 2(b) [5], where each compartment is designed to reproduce nonlinear dynamics of a membrane potential of the corresponding part of the neuron. On the other hand, our group has been developing a neural system modeling approach based on the nonlinear dynamics of an asynchronous cellular automaton, where nonlinear dynamics (especially, bifurcations) of neural systems are reproduced by the asynchronous cellular automaton with low hardware cost [6]-[9]. Our group is conceptually referring to such a hardware platform as “asynchronous bifurcation processor (ABP)” In this paper, a multi-compartment neuron model based on the concept of the ABP [9] is studied. It is shown in this paper for the first time that the model can reproduce typical propagation phenomena of membrane potentials between somas and dendrites of neurons such as forward propagations of action potentials from dendrites and resulting backward propagations of action potentials from a soma. It is also shown that the model can reproduce formations of typical potential gradients in dendrites of neurons.

2. Multi-compartment neuron model based on ABP

Fig. 1(a) shows a basic structure of a multi-compartment neuron model based on the ABP [9]. The model consists of Q > 0 compartments \{\text{C}_0, \text{C}_1, \cdots, \text{C}_{Q-1}\}, where all the compartments are assumed to be connected (i.e., there is no isolated compartment). The 0-th compartment \text{C}_0 is used as a soma compartment and the other compartments \{\text{C}_1, \cdots, \text{C}_{Q-1}\}, are used as dendrite compartments. A dendrite compartment \text{C}_i is said to be a terminal compartment if it is connected to exactly one dendrite compartment, e.g., the compartments \{\text{C}_4, \text{C}_5, \text{C}_7, \text{C}_8\} in Fig. 1(a) are terminal compartments. As shown in Fig. 1(b), each \(\text{C}_i\)-th compartment \text{C}_i can accept the following stimulation input (not necessarily).

\[ I_i(t) = \begin{cases} 1 & \text{if } t \in \{t_1^{(1)}, t_1^{(2)}, \cdots\}, \\ 0 & \text{otherwise}, \end{cases} \]

where \(t_n^{(m)}\) is the \(n\)-th spike timing (or rising edges) of the stimulation input \(I_i(t)\). As shown in Fig. 1(b), each \(\text{C}_i\)-th compartment \text{C}_i has a membrane register storing the following discrete membrane potential.

\[ V_i \in \{0, 1, \cdots, N - 1\}, \]

where the integer parameter \(N > 0\) determines the resolution of the discrete membrane potential \(V_i\). Also, each \(\text{C}_i\)-th compartment \text{C}_i has a recovery register storing the following discrete recovery variable.

\[ U_i \in \{0, 1, \cdots, M - 1\}, \]

where the integer parameter \(M > 0\) determines the resolution of the discrete membrane potential \(U_i\). (v)The \(\text{C}_i\)-th and the \(\text{j}\)-th compartments \text{C}_i and \text{C}_j are connected via discrete conductances \(G_{ij} \in \{0, 1, \cdots, L - 1\}\) and \(G_{ji} \in \{0, 1, \cdots, L - 1\}\), where the integer parameter \(L > 0\) determines the resolution of the discrete conductances \(G_{ij}\) and \(G_{ji}\). Each \(\text{C}_i\)-th compartment \text{C}_i has the following internal clocks.

\[ C_{Vi}(t) = \begin{cases} 1 & \text{if } t \in \{t_1^{(1)}, t_1^{(2)}, \cdots\}, \\ 0 & \text{otherwise}, \end{cases} \]

\[ C_{Ui}(t) = \begin{cases} 1 & \text{if } t \in \{t_1^{(1)}, t_1^{(2)}, \cdots\}, \\ 0 & \text{otherwise}, \end{cases} \]
Figure 1: (a) Structure of the multi-compartment neuron model based on the ABP. (b) Connection between the $i$-th compartment $C_i$ and the $j$-th compartment $C_j$ via the discrete conductances $G_{ij}$ and $G_{ji}$. (c) Typical time-waveforms of the $i$-th compartment. (d) Connection between the $i$-th compartment $C_i$ and the $j$-th compartment $C_j$ via the discrete conductances $G_{ij}$ and $G_{ji}$. 

- 490 -
where $t_{Gi}^{(0)}$, $t_{Gi}^{(1)}$, and $t_{Gi}^{(2)}$ represent spike timings (or rising edges) of the clocks. Let $t^r$ denote $\lim_{m \to \infty} t + \varepsilon$ (i.e., just after $t$). The internal clocks $C_{Vi}$ and $C_{Ui}$ trigger the following asynchronous transitions of the discrete states $V_i$ and $U_i$, respectively (see for example $V_i$ at $t = t_{Vi}^{(1)}$ and $U_i$ at $t = t_{Ui}^{(2)}$ in Fig. 1(c)).

\[
C_{Gi}(t) = \begin{cases} 
1 & \text{if } t \in \{t_{Gi}^{(1)}, t_{Gi}^{(2)}, \cdots\}, \\
0 & \text{otherwise},
\end{cases}
\]

where $t_{Gi}^{(0)}$, $t_{Gi}^{(1)}$, and $t_{Gi}^{(2)}$ represent spike timings (or rising edges) of the clocks.

For example, at $t = t_{Gi}^{(1)}$ in Fig. 1(d), the membrane potential $V_i$ increases by $G_{Gi}(V_j - V_i)$ since $V_j > V_i$. Also, at $t = t_{Gi}^{(2)}$, the membrane potential $V_i$ decreases by $G_{Gi}(V_i - V_j)$ since $V_i < V_j$. Fig. 3 shows reproductions of typical dendritic phenomena by the multi-compartment neuron model based on the ABP. In Fig. 3(a), weak stimulation inputs $I_4$ and $I_5$ are applied to the terminal compartments $C_4$ and $C_5$, respectively. In this case, no action potential is evoked but the membrane potentials $V_i$ of the compartments form a potential gradient. In Fig. 3(b), strong stimulation inputs $I_4$ and $I_5$ are applied to the terminal compartments $C_4$ and $C_5$, respectively. In addition, a weak background noise spike-train $n_i$ is applied to each compartment. In this case, the stimulation inputs $I_4$ and $I_5$ evoke action potentials of $V_4$ and $V_5$ in the terminal compartments $C_4$ and $C_5$, respectively. These action potentials evoke an action potential of $V_3$ in the dendritic compartment $C_3$. Repeating such dynamics, the action potential propagates to the soma compartment $C_0$. In Fig. 3(c), a weak stimulation input $I_5$ is applied to the soma compartment $C_0$. In this case, no action potential is evoked but the membrane potentials $V_i$ of the compartments form a potential gradient. In Fig. 3(d), a strong stimulation input $I_4$ is applied to the soma compartment $C_0$. In addition, a weak background noise spike-train $n_i$ is applied to each compartment. In this case, the stimulation input $I_4$ evokes an action potential of $V_0$ in the soma compartment $C_0$. This action potential evokes an action potential of $V_1$ in the dendritic compartment $C_1$. Repeating such dynamics, the action potential propagates to the terminal compartments $C_4$ and $C_5$. In Fig. 3(d), a strong stimulation input $I_4$ is applied to the terminal compartment $C_4$ and a weak background noise spike-train $n_i$ is applied to each compartment. In this case, a forward propagation induces generation of an action of the soma compartment $C_0$ and it induces a backward propagation. Note that such a backward propagation induced by a forward propagation plays important role in an STDP learning.

\[G_{ij}(V_i, V_j) = \begin{cases} 
1/8 & \text{if } -30 \leq |V_i - V_j| \leq 30, \\
0 & \text{otherwise}.
\end{cases}\]

where

\[
G_{ij}(V_i, V_j) = \begin{cases} 
1/8 & \text{if } -30 \leq |V_i - V_j| \leq 30, \\
0 & \text{otherwise}.
\end{cases}
\]

In this paper, the multi-compartment neuron model based on the ABP was studied. It was shown that the model can reproduce typical dendritic phenomena such as the forward propagation of action potentials, the backward propagation of action potentials, and the backward propagation induced by the forward propagation. It is was shown that the model can reproduce formations of the potential gradients. Future problems include development of an ABP-based multi-compartment neuron model with STDP learning capability and development if a large scale network of the ABP-based multi-compartment neuron models.
work was partially supported by JSPS KAKENHI Grant Number 15K00352.

References


Neural avalanches at the edge-of-chaos?

Karlis Kanders† and Ruedi Stoop†

†Institute of Neuroinformatics and Institute of Computational Science, University of Zurich and ETH Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland
Email: kkanders@ini.phys.ethz.ch, ruedi@ini.phys.ethz.ch

Abstract—Does the brain operate at criticality, to optimize neural computation? Literature uses different fingerprints of criticality in neural networks, leaving the relationship between them mostly unclear. Here, we compare two specific signatures of criticality, and ask whether they refer to observables at the same critical point, or to two differing phase transitions. Using a recurrent spiking neural network, we demonstrate that avalanche criticality does not necessarily lie at edge-of-chaos.

1. Introduction

In the endeavour of understanding the functioning of the brain, the hypothesis has emerged that biological neural networks might be operating at criticality [1]. The promise of this hypothesis is that at the critical point the particular details of the system’s individual elements and their interaction laws cease to be of importance. In this case, the phase transition itself dominates the behavior of the system and therefore the astounding anatomical and biophysical details of neural circuits would surrender to some very generic network properties, allowing to grasp the fundamentals of the information processing and computation in the brain. A “fingerprint” of criticality is power law distributions of the properties exhibited by local descriptors evaluated across the ensemble. Such a fingerprint was also discovered in the statistics of spontaneous activity avalanches of cortical neural tissue recorded with multi-electrode arrays [1]. While alternative explanations for power law observations have been proposed, the proponents of the avalanche criticality hypothesis attribute several advantages of this state: optimized information transmission and capacity, as well as increased flexibility of responses granted by diverse activity patterns [2, 3]. The notion of computation used in this context remained, however, rather vague, similar to the dynamical (in contrast to the topological) situation counterpart, where it has been contended that computation, in the sense of the ability of a system to transmit, store and modify information, would be optimized at edge-of-chaos [4] criticality. Links have been occasionally drawn in the literature between the edge-of-chaos and the avalanche criticality [5], however, the relationship between these two phase transitions in neural networks is far from established. Here, we examine whether avalanche criticality lies at the edge-of-chaos, characterized by a zero largest Lyapunov exponent.

2. Neural network model

The neural network model used in this study reflects the general features ascribed to cortical networks: sparsely connected and consisting of 80% excitatory and 20% inhibitory neurons, where inhibitory synapses are several times stronger than the excitatory ones. The network topology was a directed Erdős-Rényi random graph; this type of connectivity is more in line with the networks in dissociated cortical cultures [6]. The number of nodes in the network was set to \( N = 128 \), and the mean in-degree was set to \( k = 5 \), which is equal to 4% connectivity. The size of the network was chosen to satisfy a trade-off between obtaining enough statistics for the avalanche size distributions and minimizing the calculation time of the network’s Lyapunov exponents.

To assess the dynamical stability properties of the network, we need nodes that exhibit dynamics similar as possible to the membrane potential dynamics of real neurons, which can be achieved by using the Rulkov map model (cf. Eq. 3, 4 in Ref. [7] where \( u = y_n + \beta I_{syn} \) and \( \sigma_n = I_{in} \)). The parameter values for excitatory and inhibitory neurons were identical: \( \sigma = 3.6, \mu = 0.001, \sigma = 0.09, \beta = 0.133 \). Synaptic input \( I_{syn} \) was modeled by exponential decay and step-like increase upon a presynaptic spike event, \( I_{n+1}^\text{syn} = \gamma I_n^\text{syn} - \sum_j W_{\text{scale}} w_j (x_n - x_p) \), where \( \gamma \) controls the decay rate of the synaptic current, \( w_j \) is the synaptic strength between the presynaptic neuron \( j \) and the postsynaptic neuron \( i \), \( x_p \) is the reversal potential which determines whether the synapse is inhibitory or excitatory, and \( W_{\text{scale}} \) is a global scaling parameter of the synaptic weight. We used the following parameter values for excitatory (‘Ex’) and inhibitory (‘Inh’) synapses: \( x_p^\text{Ex} = 0, \gamma^\text{Ex} = 0.75, w_j^\text{Ex} = 0.6, x_p^\text{Inh} = -1.1, \gamma^\text{Inh} = 0.75, w_j^\text{Inh} = 1.8 \). Because this would pertain to self-organized criticality, which is not the goal of the present investigation, synaptic plasticity was not included in the network model.

To introduce spontaneous activity, we modelled one of the neurons to spike intrinsically by setting its parameter \( \sigma = 0.103 \), which is just above the spiking threshold. By embedding this neuron in the network, the network’s influence on spontaneous firing is much more realistic compared to if the network were solely external input-driven. Additionally, we also added a sparse, excitatory external input to all neurons in the form of independent Poisson
spike trains. The probability for a single neuron to receive an external input spike at any given iteration was \(6 \times 10^{-4}\). This type of input models randomness similar to the spontaneous neurotransmitter vesicle release.

The parameters of the network model were kept fixed except for the synaptic weight scaling parameter \(W_{\text{scale}}\) that was varied for accessing subcritical, critical and supercritical activity states. For each of the three states we ran 50 simulations and pooled the results. For each simulation, the synaptic connections were randomized. A single simulation covered \(5 \times 10^5\) iterations, and the first 5000 iterations were discarded.

3. Analysis of neuronal avalanches

Neuronal avalanches are generally defined as periods of uninterrupted neural activity, either local-field potential events or spikes, with respect to a given time binning. In our case we will analyse spike avalanches. Time is divided into bins of length \(\Delta t\) and an avalanche is a sequence of bins that each features at least one spike, preceded and followed by at least one bin without any spikes. The lifetime of the avalanche \(T\) is the number of bins in the sequence. The size of the avalanche \(S\) is the sum of the spikes in the sequence. A popular decision in the experimental studies is to use a temporal bin size equal to the average time between two subsequent events across all of the electrodes: the inter-event interval \(IEI_{\text{avg}}\) [1, 8, 9, 10]. In order to put our study in the same context as the experimental investigations of the avalanche criticality, we will follow the established approach of defining neural avalanches by using a binning size of \(\Delta t = IEI_{\text{avg}}\) [1].

Power law distributions can be caused by several different mechanisms and not necessarily by a phase transition, which necessitates additional tests to confirm that the network is really at criticality. An important test for the scale-free property of the avalanches is the universal scaling of avalanche shapes [10]. The avalanche shape of length \(T\) is defined as the temporal profile of an avalanche, i.e., the number of spikes over time \(V(T, t)\). The critical point is characterized by power laws in many variables and these relationships give rise to the fractal structure of avalanches: the average shapes are similar over different time scales and they collapse to one universal shape after rescaling. In contrast, subcritical or supercritical avalanche shapes should not collapse. We let \(\langle V(T, t/T) \rangle\) be the average temporal evolution of the size of an avalanche with a normalized duration \(t/T\) and rescale the avalanches to \(V(t/T) = T^{1-\gamma} \langle V(T, t/T) \rangle\), where \(V(t/T)\) is the universal scaling function, i.e., the characteristic shape of all avalanches. The critical exponent \(\gamma\) can be obtained from \(\langle S(T) \rangle \propto T^\gamma\), where \(\langle S(T) \rangle\) is the mean size of avalanches as a function of the duration \(T\), which can be easily estimated from the simulation results.

Figure 1: Avalanche size \(S\) distributions: (a) subcritical, (b) critical, (c) supercritical networks. The critical network shows a power law behavior for \(S \geq 6\), with \(\alpha \approx 2.4\) (fitted using the maximum-likelihood estimation; goodness of fit evaluated using the Kolmogorov-Smirnov distance with 1000 synthetic samples [11], p-value = 0.52). The decay of the subcritical distribution can be fitted by an exponential with exponent \(\lambda \approx 0.21\) (p-value = 0.26).

4. Results and discussion

By globally increasing the synaptic strength of the connections with the parameter \(W_{\text{scale}}\), we observed an overall increase in network activity and an evolution of the topological network state from subcritical, to critical, to supercritical (Fig. 1). The values of \(W_{\text{scale}}\) corresponding to subcritical, critical and supercritical networks were 0.13, 0.139 and 0.15, respectively. The mean \(IEI_{\text{avg}} \pm SD\) was 110 ± 8 (subcritical network), 48 ± 6 (critical network) and 8 ± 2 (supercritical network) iterations. Changing \(W_{\text{scale}}\) appeared to have a similar effect to changing the levels of excitation in biological experiments by using pharmacological agents that alter the efficiency of neurotransmitter receptors [3]. The avalanche size distribution of the critical network follows a power law with the exponent \(\alpha \approx 2.4\) (Fig. 1(b)).
with a power law noise cut-off at about $S \approx 100$. This is expected because the network is a finite system of the size $N = 128$ and in most of the avalanches a single Rulkov neuron fires only once. In the subcritical case, the avalanches were smaller and their size decayed exponentially, while in the supercritical case there was an increased number of large avalanches signified by the hump at the end of the distribution. If the synaptic strength were to be increased further, the hump would become even more prominent. Similar metamorphosis of the distribution shape was also observed for avalanche lifetimes. At criticality, the lifetime distribution could be fitted with a power law for $T \geq 10$, with an exponent $\tau = 2.9$.

The critical exponent $\alpha = 2.4$ of our avalanche size distribution is different from the slope $\alpha = 1.5$ measured in the original experiments [1]. However, in subsequent reports of critical spike avalanches the exponents have varied substantially, in the range from 1.5 to 2.1 [8, 9]. Some researchers explicitly reject the idea of a universal exponent (such as $\alpha = 1.5$) because of insufficient experimental and theoretical basis [12]. Our critical avalanche size distribution is rather similar to that of dissociated rat cortical neurons reported in Ref. [9], which had $\alpha = 2.1$ at the tail of the distribution and another scaling regime for very small avalanches.

The critical network showed a rather noisy collapse of the avalanche shapes of duration $T \geq 20$ (Fig. 2(a)), but overall the avalanche shapes were more self-similar than in the case of the supercritical network (Fig. 2(b)). The scaling could not be assessed for the subcritical network, as the maximum lifetime of the subcritical avalanches was only $T_{\text{max}} = 26$. A collapse similarly noisy to ours was reported for avalanches recorded in dissociated neuron cultures [10], which have a topology similar to our network model [6].

As the final test, we examined the relation among critical exponents $\frac{\tau-1}{2} = \gamma$ [10]. The critical exponents of avalanche lifetime distribution ($\tau = 2.9$), avalanche size distribution ($\alpha = 2.4$), and the function of the mean avalanche size depending on the lifetime ($\gamma = 1.37$) fulfill this relation. Taken together: power law distributions, the similarity of avalanche shapes and an excellent fulfillment of the fundamental relation between critical exponents, strongly suggests that our ‘critically tuned’ network is indeed critical.

The largest Lyapunov exponents for the three network states, calculated using the network’s Jacobian matrix evaluated at points along the trajectory of the state vector [13], were, however, all positive and practically the same for subcritical and supercritical networks and slightly smaller for the supercritical network: $\lambda_1 \approx 18 \ s^{-1}$ for subcritical and critical networks and $\lambda_1 \approx 16.5 \ s^{-1}$ for the supercritical network (applying the time rescaling of [7] with one iteration accounting for about 0.5 ms).

The Lyapunov spectra in the three cases provided more insight by showing that the number of positive Lyapunov exponents increased with coupling strength. The upper bound of the Kolmogorov-Sinai entropy $H = \sum_{i=0}^{\infty} \lambda_i$ in our networks increased with a higher coupling strength and was the highest for the supercritical network: 28 $s^{-1}$ (subcritical), 46 $s^{-1}$ (critical), 88 $s^{-1}$ (supercritical). Although the supercritical network has a slightly smaller largest Lyapunov exponent, it loses the information about the past states at the fastest rate.

As chaotic dynamics could be a collective effect of the network interactions or arise simply because the nodes themselves have chaotic dynamics, we measured the largest Lyapunov exponent of the intrinsically spiking neuron and found it to be positive ($\lambda_1 = 20 \ s^{-1}$), in the absence of network input. In the presence of external input, the neuron was occasionally silenced, a behavior that can be also observed in Class II neurons. As a result, the neuron’s largest Lyapunov exponent decreased to $\lambda_1 \approx 18 \ s^{-1}$ which is in agreement with $\lambda_1$ of our networks. This suggests that the largest Lyapunov exponent of the network might be capturing the dynamics of the intrinsically spiking neuron. In the subcritical and critical cases there were 3–4 other positive Lyapunov exponents, which is close to the number of neurons that receive inputs from the intrinsically spiking neuron. Therefore, the source of chaos in our networks resides in the single neuron dynamics; the increase of coupling strength made the chaos more intensive because it allowed more neurons to spike.
5. Conclusion

Our network showed chaotic dynamics, for all choices of the synaptic weights. The main result of the present study is that we did not observe a coincidence of avalanche and edge-to-chaos criticality in our network type. This suggests that in neural networks with non-trivial node dynamics these are two separate phase transitions. Both phase transitions have been occasionally mentioned in the same context in the literature; our demonstration also elucidated that any computational benefits of one criticality cannot be directly translated to the other, and that in this regard the internal neuronal processes appear to play a decisive role.

References


Novel insights into cochlear information processing

Ruedi Stoop, Karlis Kanders, Leonardo Novelli, Florian Gomez†‡

† Institute of Neuroinformatics and ‡ Institute of Computational Science, University and ETH Zürich
Irchel Campus, Winterturterstr. 190, 8057 Zürich, Switzerland
Email: ruedi@ini.phys.ethz.ch

Abstract—Already Helmholtz profoundly addressed the question how the nonlinearity of the human hearing sensor, the cochlea, might shape human sound perception. At his time, research was, however, obstructed by the lack of experimental data regarding the amplification properties of the inner ear. In the meantime, accurate measuring methods have permitted the comparison of models of the hearing sensor with empirical data, leading to a strong revival of the interest into Helmholtz’s original research questions. In our paper, we describe some recent theoretical and modeling advances in the understanding of the nature of human pitch perception. We reveal a number of to date unexplained human auditory percept effects to be direct consequences of the nonlinear properties of the mammalian hearing sensor. Our insights also demonstrate, as a by-note, the limitations of the present reverse engineering approach towards cochlear implants.

1. Introduction

Audition has a long history, starting with the ancient Greeks who stipulated that inside the ear there should be air, as matter only communicates with matter of the same type. Only in the 1800s was it discovered by the physiologist Cotugno that in fact, the inner ear, the cochlea, is filled with perilymph comparable to that of plasma and cerebrospinal fluid, a salty fluid and endolymph that resembles more intracellular fluid [1]. A similar fate is related to the understanding of the information processing within the cochlea. While it was early realized by Helmholtz [2] that nonlinear processes play a major role, probably driven by von Bekesy’s experiments and his profound understanding of the passive (i.e., essentially dead) cochlea, the picture prevailed that information processing in the cochlea would be very similar to a Fourier-analysis tool. Even after the active processes in the cochlea were postulated by Gould and finally corroborated by the detection of otoacoustic emissions by Kemp, this view basically remained, although it was widely known that nonlinear active amplification must be expected to generate strong deviations from a linear model. Combination tones (‘CT’) are a major effect of the nonlinearity of the physics underpinning amplification in the mammalian hearing process. They have a profound effect on different aspects of how we perceive sounds, one main manifestation is the perception of pitch. It was probably the consequence of the negation of the important role of nonlinearity that CT were also termed ‘distortion products’, with the erroneous assumption that CT would only have a noticeable role in hearing at very strong input amplitudes. In fact, their role is often more important at small amplitudes. Besides this more local effect (CT travel down the cochlear duct just as directly input-generated sounds do, and they are also locally amplified like the latter) a second, more global manifestation of the nonlinearity located in the cochlea’s amplification properties, exists. This effect is the behavior of the mammalian hearing threshold that is known to be frequency-dependent. We will show that this is not because the amplifiers towards the ends of the accessible frequency interval are less efficient, but that this happens as the consequence of a more global behavior emerging from local nonlinearities and their couplings.

2. Combination-Tone Laws

For nonlinear systems, the superposition principle does not hold. This implies that, in general, the response to the sum of two inputs is not the sum of the individual responses to each input. In the cochlea, this manifests in a variety of effects, the most prominent one being the generation of CT within the cochlea. CT were already known to 18th century musicians Georg Sorge and Giuseppe Tartini, and can easily be heard also by non-trained listeners. CT are by-products that, however, are not filtered out on the way to the brain. Instead, they propagate along the cochlear duct, get amplified, and interact with other frequencies to create additional CT. They thus were detected at all stages of the auditory pathway, e.g., in the inner hair cell response, in the auditory nerve, or in the inferior colliculus. From this, it follows that the cochlear activation profiles often become highly nontrivial, despite the simple inputs used.

Single Hopf Oscillator: First, we will present the basic principles of combination tone generation using the example of a single Hopf oscillator. A detailed analysis however shows that only the full Hopf cochlea model can exhibit combination tone responses comparable to the ones observed in biological and psychoacoustic experiments. The following analysis exhibits in detail how a single Hopf oscillator generates CT. Upon a stimulation with two frequencies, as a product of the nonlinearity in the Hopf equation, a well-defined set of CT is generated (‘cubic CT’). Cubic CT show, as a function of their order, exponentially decaying amplitudes, a fact that has been mentioned in the
context of early psychoacoustic studies as well as more recent investigations of cochlear mechanics (see also [3]).

From the \( \omega_{ib} \)-rescaled Hopf equation of a section [3, 4]

\[
\frac{dz}{dt} = (\mu + i) \omega_{ib} z - \omega_{ib} z^2 - \omega_{ib} F(t),
\]

where \( \omega_{ib} \) is the characteristic frequency of the oscillator, \( F(t) \) is the (complex) external forcing and \( \mu \) is the bifurcation parameter, a harmonic-two-tone forcing \( F(t) = F_1 e^{i \omega t} + F_2 e^{i \omega t} \), with \( \omega_1 = k \omega_0 \) and \( \omega_2 = (k + 1) \omega_0 \) so that all CT are multiples of \( \omega_0 \), generates a response that can be expanded in a Fourier series \( z(t) = \sum_i a_i e^{i \omega_i t} \), with complex coefficients \( a_i \) (i.e. they include a phase). By inserting the expansion of \( z(t) \) for a frequency \( \omega_1 = k \omega_0 \) into Eq. (1), we obtain

\[
(\mu \omega_i - \omega_{ib}) a_i + c.i.t. = -\omega_{ib} F_i,
\]

where c.i.t. denote cubic interaction terms (\( \propto \omega_{ib} a_i a_j a_k \), where \( k' + k'' = k + \mu' \)).

We now may calculate the responses at \( \omega_1, \omega_2 \) and all combination tones, if we assume low to moderate sound levels and an \( \omega_{ib} \) close to the forcing frequencies. Inserting the expansion of \( z(t) \) into the Hopf equation, to lowest order in \( \omega_1 \) and similarly for \( \omega_2 \) we obtain

\[
i(\omega_1 - \omega_0) a_{CT1} = (\mu + i) \omega_{ib} a_{CT1} - \omega_{ib} a_1 a_1^*.
\]

For higher sound levels, to the r.h.s. the cubic interaction terms

\[
-2 \omega_{ib} a_{CT1} |a_1|^2 - 2 \omega_{ib} a_{CT1} |a_1|^2
\]

would have to be added. This, however, retains the equation to be linear in \( a_{CT1} \), since the \( |a_{CT1}|^2 a_{CT1} \)-term is of higher order. The next combination tone CT2 follows to lowest order (\( \propto (a_1^3)^3 \)) from

\[
i(\omega_1 - 2 \omega_0) a_{CT2} = (\mu + i) \omega_{ib} a_{CT2} - 2 \omega_{ib} a_{CT1} a_1 a_2^*.
\]

For the third combination tone CT3 at \( \omega_1 - 3 \omega_0 \), we proceed correspondingly and get to lowest order

\[
i(\omega_1 - 3 \omega_0) a_{CT3} = (\mu + i) \omega_{ib} a_{CT3} - 2 \omega_{ib} a_{CT2} a_1^2 - \omega_{ib} a_{CT1} a_2^2.
\]

Here, the last two terms are of the same order (\( \propto (a_1^3)^3 \)). Using Eqs. (5) and (9), for \( \omega_{ib} \approx \omega_0 \) and \( \mu = 0 \) we obtain \( a_{CT3} \approx \omega_{ib} a_1 a_2^* / |\omega_0| \), which is exactly of the same form as Eq. (9). Handling the interaction terms of lowest order carefully shows that the same law holds for all subsequent combination tones CT4, CT5, ..., which leads to an approximate exponential decay of CT amplitudes as

\[
a_{CTk} \approx \frac{\omega_{ib} a_1 a_2^*}{|\omega_0|} a_{CT(k-1)} = \kappa a_{CT(k-1)},
\]

with \( \kappa := (\omega_{ib} a_1 a_2^*)/(i \omega_0) \). Approximate exponential decays emerge off-bifurcation as well (e.g. \( \mu = -0.1 \)), and not exactly at resonance \( \omega_{ib} = \omega_0 \), see below.

Two examples confirm these results. For the first we choose a two-tone stimulus of \( \omega_1 = 2 \pi \cdot 2000 \text{ rad/s} \) and \( \omega_2 = 2 \pi \cdot 2200 \text{ rad/s} \) with amplitudes \( F_1 = F_2 = 0.01 \). In Fig. 1a) we then compare the numerical integration of Eq. (1) with our above-derived analytical calculations for \( \omega_{ib} = \omega_0 \) and \( \mu = 0 \): For the theoretical approach, \( a_1 \) and \( a_2 \) follow from solving Eqs. (3) and (4) simultaneously, which yields \(-14.1 \text{ and } -22.3 \text{ dB respectively, } a_{CT3} \) follows from Eq. (5) together with Eq. (6), yielding a response of \( |a_{CT3}| \approx 3.36 \text{ dB} \). Using this, we obtain (to lowest order) from Eq. (7) \( a_{CT2} \) and from Eq. (10) \( a_{CT3} \) (and correspondingly the other CT). For the numerical integration, we chose a sample rate \( S R = 80 \text{ kHz} \) and a fourth-order Runge-Kutta scheme with integration step \( h = S R^{-1} \), whereupon we observe an excellent agreement. For the second example, we chose \( \omega_2/\omega_1 = 1.05 \) and \( \omega_{ib} = 2 \omega_1 - \omega_2 \), which corresponds to the biological experiment of [6]. We use \( \omega = 2 \pi \cdot 2000 \text{ rad/s} \) and \( \omega = 2 \pi \cdot 2100 \text{ rad/s} \) for simplicity (\( \Delta f := (\omega_2 - \omega_1)/2 \pi = 100 \text{ Hz} \)), and amplitudes \( |F| = 0.01 \text{ (40 dB, moderate to high sound level) \). Fig. 1 shows the obtained response b) for a biologically reasonable value of \( \mu = -0.1 \) and c) at criticality, \( \mu = 0 \). Both settings produce exponential CT amplitude decays, but with significantly too high decay exponents for corresponding sound pressure levels (Biology: 5-6 dB/\( \Delta f \) for the lower CT [6]). For the whole set of biological measurements from 30 to 80 dB SPL made in [6], a single Hopf oscillator underestimates CT levels substantially. This is, however, not the case for the compound Hopf cochlea, as will be exhibited below.

**Full Hopf Cochlea:** In biology, CT of frequencies lower than stimulus propagate down the cochlea until the waves are amplified and stopped where their frequency matches the characteristic frequency \( \omega_{ib} \). This situation that differs from the single Hopf element case, leads to an asymmetric (low-pass) and generally slower CT decay. We first exhibit the response corresponding to the first cubic combination.
tone ‘CT1’, i.e. the $2f_1 - f_2$-tone. For this, we focus on 
a location with $\omega_{ch} = \omega_{CT1}$, to see how strong a two-tone signal from two single pure tones of equal strength would 
be required to be to generate the same effect as a direct stimulation by a tone of frequency $f_{ch}$ for a single Hopf oscillator, and for a compound cochlea as well. We then 
compare these results to the biological measurements. The conventional quantification of this difference is the ‘relative 
CT1-tone strength’, that evaluates at a characteristic place 
$x_{ch}$ how much stronger a two-tone input (the two inputs of equal strength) having a CT1 at $\omega_{ch}$ would be needed 
to generate the same response as a pure tone with $\omega_{ch}$ [6]. The 
horizontal distances between the black and the green lines in Fig. 2a, b) exhibit this measure. The obtained 
results show that CT1-amplitudes depend in a nontrivial man-
ner on stimulation level, and also on whether we consider 
a Hopf amplifier alone or a section within the compound 
cochlea. Only the compound cochlea reproduces the effect 
correctly, and is also consistent with measurements from 
the apical part of the cochlea [7].

We now focus on the full set of cubic combination tones. These results (not shown) demonstrate that the CT are 
stronger for higher input levels; nevertheless, the first CT at frequency below the stimulation frequencies $f_1$ and $f_2$ 
(i.e. at frequency $2f_1 - f_2$) is clearly visible even for an 
input level as low as 30 dB SPL (–84 dB). CT are thus not 
only relevant at high sound levels. As we will see below,

Figure 1: CT in a single Hopf oscillator. a) Fourier transform of the response of a single Hopf oscillator ($\omega_{ch} = 2\pi \cdot 2000$ rad/s, $\mu = 0$) when subject to two-tone forcing ($\omega_1 = 2\pi \cdot 2000$ rad/s and $\omega_2 = 2\pi \cdot 2200$ rad/s with 
amplitudes $F_1 = F_2 = 0.01$); spectrum: numerical integration, crosses: analytical results; exponential decay factor 
$|\kappa| = \frac{\omega_{ch}^2}{\omega_1 \omega_2}$ from Eq. (11). b), c) Same experiment as in 
a), with $\omega_{ch} = 2\pi \cdot 1900$ rad/s, $\omega_{1,2} = 2\pi \cdot (2000, 2100)$ rad/s 
and amplitudes $F_1 = F_2 = 0.01$, where b) $\mu = -0.1$ and c) $\mu = 0$ (oscillator at criticality).

Figure 2: The behavior of the first cubic CT in a single 
Hopf element, in the compound cochlea and in biology. 
a,b) Response amplitude to a pure tone of frequency $f_{ch}$ 
(black), response to a two-tone input (equal strength of 
components) with $f_{ch} = f_{CT1}$ (green). The difference (ar-
rows) is the ‘relative strength of CT1’. a) Single Hopf am-
plifier (no fluid comprised), and b) cochlea section 6 where 
$f_{ch} = f_{CT1}$. c) Relative strength of CT1 for two $f_2/f_1$-
frequency ratios. Red: cochlea section 6, black: biologi-
dal data [6] ($f_{ch} = 9000$ Hz). The blue arrows in b) and c) 
describe the same experimental result.
(including $f_1, f_2$) have been dissipated, and the remaining frequencies are grouped around the local characteristic frequency $f_{ch} = 1.31$ kHz. Towards the end of the cochlea, also the remaining lower CT vanish gradually.

3. Hearing Threshold and Generalized Hopf Cochlea

We then looked at the hearing threshold curve of the Hopf cochlea and compared it with the psychoacoustic hearing threshold curves of human subjects in age groups 10-21 and 56-65 years [11], where our Hopf cochlea comprised 23 sections ($\mu = -0.25$) with characteristic frequencies covering the frequency interval from 311 to 14.08 kHz. Input signals were pure tones (no higher harmonics). A second surprising manifestation of the, this time inherently global, nonlinear character of the hearing sensor is that the hearing threshold curves that characterize the finest sound that elicits a neuronal response obtains a measurable response are very well reproduced by the Hopf cochlea. This observation is based on the conversion of the cochlea measurements to sound pressure units, where a cochlear amplification threshold of -50 dB corresponds to the psychoacoustic hearing threshold. Declaring a section excited, if the response reaches above this level, without any tuning an extremely good approximation of the u-formed psychoacoustical hearing threshold is obtained. Since this emerges despite this uniform tuning of the compound cochlea, this effect is of a global nonlinear nature of the hearing device, observed even for stimulations that are fully based on pure tones (no higher harmonics are present in the input signal).

Already v. Helmholtz [2] had speculated on the origins of the hearing nonlinearities; he saw it in the eardrum nonlinearities or the mechanical impedance-matching middle ear, an ‘error’ that was finally corrected by v. Bekesy’s measurements [9] (a concise account of the time around v. Bekesy’s discovery is found in Ref. [10]). Whereas Hopf systems with a cubic nonlinearity ($\gamma = 2$, below) are consistent with biological experiments, we may for completeness also study a generalization to values of $\gamma$ different from two, by considering in Eq. (1) instead of the term $\omega_{ch} |z|^2 z$ a term $\omega_{ch} |z|^2 z^\gamma$, with variable exponent $\gamma$. While the decay of the combination tones is a property of the Hopf system, depending, in particular, on the exponent $\gamma$, systems with $\gamma \neq 2$ can generate the same effects, but show different compressive nonlinearities and different decay laws of the combination tones. Following the experiments by Smoorenburg [12], the pitch of a sound is extracted in the cochlea in a neighborhood where the lowest audible CT is found (the pitch itself then is the residue pitch of the spectrum obtained at this location [8]). A generalized Hopf system in particular leads to the response $| F |^\frac{1}{\gamma'}$. This also alters the CT decay behavior and will hence also modify the pitch perception, as a function of $\gamma$.

Acknowledgment: The authors acknowledge supporting grants SNF (200020-147010) and ETHZ (ETH-37 152).

Figure 3: CT decay, generalized Hopf cochlea, $\gamma = 1, 2, 3$.

References

Forecasting Correlation Structures

Martin Schüle†, Thomas Ott† and Peter Schwendner‡

†Institute for Applied Simulation, Zurich University of Applied Sciences, Switzerland
‡ZHAW School of Management and Law, Zurich University of Applied Sciences, Switzerland
Email: scll@zhaw.ch

Abstract—Often the signature of a complex system is a couple of empirically found time series. As the exact processes generating these series are often unknown, one contents oneself with mere data analysis, i.e., an analysis of the statistical features of the time series. Beyond the individual statistical characteristics of the time series, a key tool to investigate the structural behaviour of the complex system is considering the correlation structure, i.e. the system of pairwise correlations between the time series.

Lacking a parametric model, the mere data analysis, i.e., the analysis of the correlation structure, may shed light only on the past behaviour of the time series. As the correlation coefficients are calculated from a given data set, there is a priori no way to forecast the future correlated behaviour of the system. However, if there are certain consistent patterns to be found in the analyzed correlation structure, the past structure assumingly allows to forecast, at least for short time periods, the joint behaviour of the system.

Often, the correlation structure exhibits some sort of dynamics which can be visualized by calculating the moving-window correlation matrices \( C(T_n) \) for periods \( T_n \). In many examples of complex system, the dynamic correlation structure then shows certain patterns that form and cluster and dissipate again over time. It is precisely these patterns that are of interest in the analysis and possibly forecasting of the correlation structure.

As an example consider a data set of daily changes in 25 different asset prices over the period June 1998 to August 2015. The data set contains 25 major assets from 4 different sectors, namely bonds, commodities, currencies and equity. Figure 1 shows the correlation matrix calculated over a period of 150 days at different times, visualized as a heat map. Clearly, block structures emerge and change over time, reflecting the relationships between the different asset classes.

We present the following approach to forecast correlation matrices. First, the eigenmodes of the correlation matrices over time are analyzed for principal oscillation patterns [1]. That is, the correlation matrices are decomposed by eigendecomposition and the eigenvalues and -vectors analyzed for characteristic oscillations patterns in time. The consistency of the oscillations patterns is assessed by a statistical bootstrap analysis and the ensuing economic interpretation. Consistent eigenmode oscillations are then forecasted by multivariate autoregressive and general trend-following and mean-reversion models [2].

By reversing the eigendecomposition, the eigenmodes oscillation forecasts allow inferring the correlation matrices. The inferred correlation matrices need further be regularized to meet the conditions of proper correlations matrices such as positive definiteness. Finally, the goodness of the forecasts are statistically assessed and the correlation forecasts interpreted in view of the economic theory.

Note that although a data set of financial time series was used to introduce and assess our correlation structure forecast method, the general method can be of use in any field with the need of analyzing correlation structures, such as e.g. in climate research [3].

References

Modeling of Human Spontaneous Eyeblinks

Ryota Nomura†; ‡ and Tohru Ikeguchi‡

†Faculty of Management Science, Tokyo University of Science
6–3–1 Niijuku, Katsushika, Tokyo 125-8585, Japan
‡Faculty of Education, The University of Tokyo
7–3–1 Hongo, Bunkyo, Tokyo 113-0033, Japan
Email: nomura@hisenkei.net, tohru@rs.tus.ac.jp

Abstract—Recent years, researchers have reported that human spontaneous eyeblinks could synchronize in accordance with external visual stimuli. In this paper, to model human eyeblinks, we provided a leaky intergrate-and-fire model with a variable threshold that represents underlying slow oscillation. To test the dynamics of eyeblink likelihood, numerical simulations were performed by changing parameters of the variable threshold. The results demonstrated positively skewed distributions of inter-blink intervals, which were reported in previous behavioral researches. Possible mechanisms of a variety of eyeblink-related phenomenon were discussed.

1. Introduction

Some researchers have reported that viewers likely to blink at implicit breaks in storytelling performances [1] and in video stories [2]. However, the emerging dynamics of eyeblink synchronization among viewers remain as a problem. One manner to approach this problem is to explore the dynamics of eyeblink synchronization by using numerical simulation. Thus, in the present study, we aim to make a differential equation model of human spontaneous eyeblinks.

Human spontaneously blinks 20 – 30 times per minute [3]. This is approximately 10 times of frequency that is enough to keep humidity of eye surfaces. As spontaneous eyeblinks, human exhibits not only periodical blinking but also quick repeated blinking during a few seconds (i.e., eyeblink bursts [4]). Numerous behavioral researches demonstrate that these blinkings correspond to attentional shifts in cognitive tasks [2].

The recent neurological studies have found that default mode network (DMN) is activated during eyeblinks [5]. The DMN is interpreted to relate to engagement in self-referential mental activities. Hence, some researchers suppose that human would blink in order to withdraw attention back from involving external targets, and then engage in inner processing [5].

These findings suggest that human blinks derive from both endogenous and exogenous factors. That is, human eyeblinks are regulated in accordance with external inputs concerning involved tasks, while at the same time, dominated by monitoring the surrounding environment in rest-
steps). Therefore, the intensity of the external input \( I \) is set to \( s = 0.04 \), taking into account stochastic distribution of \( I \). If the decay parameter \( c \) becomes large in proportion to \( s \), \( L \) would never reach the threshold. Thus, \( c = 0.023 \) was relatively selected against to \( s \).

When \( \tau = 400 \), the frequency is 0.5 [Hz]. By taking the limit \( \tau \to \infty \), \( \omega \) approaches 0. In this limit, the threshold takes constant \( f(t) = a \). The case that \( k = 0.10 \) and \( k = 0.20 \) corresponds to 10.0 % and 20.0 % of the threshold, respectively.

Table 1: Statistical values of simulated and observed [8] inter-blink intervals [sec]

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>Median</th>
<th>Mean</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation</td>
<td>0.590</td>
<td>1.820</td>
<td>3.268</td>
<td>21.420</td>
</tr>
<tr>
<td>Observation</td>
<td>–</td>
<td>1.76 ± 2.4</td>
<td>4.3 ± 0.8</td>
<td>–</td>
</tr>
</tbody>
</table>

3.2. Actual human spontaneous eyeblinks

An observational study [8] has reported a variety kinds of statistical values regarding human spontaneous blinking. In the study, blinking of 10 resting subjects without any eye abnormality were observed for 24 minutes. According to Ref. [8], human spontaneous blinking ratio was averagely 17.6 ± 2.4 blinks/minutes and the mean inter-blink intervals (IBI) was 4.3 ± 0.8 [sec]. The distribution is positively skewed, and thus the median of IBIs 2.7 ± 0.5 [sec] was lower than mean IBI. At the same time, logarithm of IBIs probability density and IBIs showed a power law when IBI > 1.025. The exponent \( \alpha \) for the power law distribution of scaling, calculated across all subjects, was \(-1.24\). On the other hand, some participants [3] exhibit a bimodal distribution with modes approximately 0.5 [sec] and 5.0 [sec]. The short inter-blink periods reflect eyeblink burst.

---

Figure 1: Behaviors variability of \( L \) in accordance with threshold parameters \( \tau \) and \( k \).

The satisfactory model must represent both periodicity and bursts of blinking.

---

Figure 2: Variation of distributions of inter-blink interval in accordance with \( \tau \) and \( k \).
3.3. Frequency and distribution of inter-blink intervals

First, we conducted a simulation with using a leaky integrate-and-fire model with a constant threshold \( f(t) = 1 \), where \( \tau \to \infty \) thus \( \omega = 0 \). Figure 1(a) shows the typical behaviors of \( L \) in these cases (only first 6,000 steps were plotted). The simulated inter-blink intervals prolonged owing to decay of current \( L \).

Second, we compared the results of simulations in order to test how \( L \) behaved in accordance with threshold parameters length of period \( \tau \) and amplitude \( k \). Figure 1 demonstrated the variation of \( L \)'s fluctuations under respective conditions. The bimodal distributions were found when the threshold is variable, both for \( k = 0.10 \) or \( k = 0.20 \). The outline of the distributions was shown in Figure 2.

We finally exerted a simulation 1000 times under the condition that \( \tau \to \infty \), \( \sigma^2 = 0.0015 \), and 24,000 steps(240.0[sec]) in order to gain a distribution of stationary blinking. In each performance, the initial values of threshold function were set differently, while all of the initial value of \( L \) were set to 0.0. Table 1 shows the statistical values of IBIs obtained from simulations and observations. The distributions of inter-blink intervals were shown in Figure 3. Figure 3(a) is simulated distribution and Figure 3(c) is an actual observational distribution (supplement data in Ref. [10]) obtained from 14 participants who were viewing videotaped storytelling performances for duration of approximately 50 minutes.

4. Discussion

4.1. A model of human spontaneous eyeblink

According to existing studies [3], [4], [11], inter-blink intervals of spontaneous blinks typically show positively skewed distribution. Moreover, the distribution approximate to logarithmic normal distribution when sufficient size of samples was collected [11].
As the result of simulations, IBIs demonstrated positively skewed distribution (Fig. 3(a)). The logarithmic transformed data (Fig. 3(b)) was near to normal distribution, and thus inter-blink intervals reproduced by the proposed model would be logarithmic normal distribution. Logarithm of IBIs probability density and distribution of IBIs showed a power law when $IBI > 1.025$. The exponent for the power law of scaling was $3.68$, which was keener compared to $1.24$ calculated using observations. Regarding IBIs, the ratio of $3.268$ per minutes was a little smaller than usual observations of human spontaneous eyeblinks [3]. However, the model in this study reproduced the positively skewed and long-tailed distribution, which characterizes the human spontaneous eyeblinks. Depending on the leak term weighted $c$, the eyeblink likelihood $L$ fluctuated near the threshold function. The IBIs distribute in a long-tailed way because the eyeblink likelihood $L$ would delay to reach the threshold when the threshold takes a positive amplitude, even if the input $I$ intermittently increase $L$.

### 4.2. Variation of inter-blink intervals due to variable threshold

The results suggested that length of period and amplitude parameters of the threshold influence on inequality of inter-blink intervals. In particular, the amplitude coefficient $k$ of threshold function provided two peaks in the distribution of inter-blink intervals. Eyeblink bursts could be explained by the variable threshold. The model in the current study could demonstrate both periodicity of blinking and eyeblink bursts by changing parameter of the threshold. Some eyeblink-related phenomena could be explained by this variable threshold. For instance, decrease of eyeblink ratio of patients with Parkinson’s disease which is related to decreasing level of a dopamine [9].

![Figure 4: Log inter-blink interval probability of simulated data as a function of the log inter-blink intervals.](image)

References


Numerical Analysis on Synchronization of Four Metronomes

Yusuke Shiomiy, Yutaka Shimada†, Kantaro Fujiwara‡, Tohru Ikeguchi‡

†Department of Management Science, Graduate School of Engineering, Tokyo University of Science
‡Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6-3-1 Nijuku, Katsushika, Tokyo 125-8585, Japan
Email: shiomi@hisenkei.net

Abstract—We numerically investigate behavior of four weakly coupled metronomes using equations of motion of the metronomes. In the numerical simulations, the parameter values of the equations of motion are experimentally decided from the experimental apparatus, where the four metronomes are put on a board hung by strings. As a result, we found that synchronization of metronomes depend on frequency and initial angles. In addition, we also found the importance of the individual difference between metronomes, which becomes clear by the numerical simulations.

1. Introduction

In our world, various rhythms exist, for example seasonal transitions, neuronal activities, and collective behavior of animals and insects[1][2]. In these real systems, even though each individual has its own rhythm, it is widely acknowledged that their rhythms synchronize by mutual couplings or external forces. We can also observe the synchronization by using weakly coupled metronomes[3][4][5]. For example, if we put a number of metronomes on a board hung by strings, the metronomes eventually synchronize.

In this paper, we numerically investigate the synchronization phenomenon of metronomes by using a mathematical model of the motion of four metronomes which are put on a board hung by strings. When we conduct numerical simulations, we used the parameter values in the mathematical model estimated from the handmade experimental apparatus. As a result, the time necessary for the in-phase synchronization becomes shorter when the frequency of the board becomes larger.

2. Mathematical model

Figure 1 shows a handmade experimental apparatus. The experimental apparatus is comprised of four metronomes put on a wooden board hung by four strings. The equation of motion of the experimental apparatus shown in Fig. 1 is approximately described as follows:

\[ \ddot{\theta}_i = -2\xi_i \omega_i \dot{\theta}_i - \omega_i^2 \sin \theta_i - (\omega_f / \omega_p)^2 \dot{\theta}_p \cos \theta_i, \]  \hspace{1cm} (1)

\[ \alpha \ddot{\theta}_p = -2\xi_p \omega_p \dot{\theta}_p - \alpha \omega_p^2 \theta_p - \gamma N \sum_{i=1}^{N} \dot{\theta}_i (\theta_p \sin \theta_i + \cos \theta_i) \]
\[ -\omega_p \gamma N \sum_{i=1}^{N} \dot{\theta}_i^2 (\theta_p \cos \theta_i - \sin \theta_i), \]  \hspace{1cm} (2)

where \( \theta \) [rad] is an angle, \( \omega \) [rad/s] is a natural angular frequency, \( \xi \) is a damping ratio, \( \alpha = (1 + N \gamma) \), \( \gamma \) is a ratio of mass of a metronome to that of the board, \( \eta \) is a ratio of the length of the pendulum attached the metronome to the length of the string. The subscript \( i \) of \( \theta, \omega \), and \( \xi \) corresponds to the \( i \)th metronome, and \( p \) corresponds to the board. In the numerical simulations, we determined the parameter values in Eqs.(1) and (2) from the experimental apparatus shown in Fig. 1. The oscillation frequency \( \omega_f \) of the \( i \)th metronome is estimated from a metronome called ‘Lupina’ produced by Nikko Seiki Co., Ltd [4]. Even if we set \( \omega_f (i = 1, 2, 3, 4) \) to 1.4 [Hz] (168[bpm]) by adjusting the sliding weight of the pendulum in Lupina, there exist individual differences between the metronomes. Then, oscillation frequencies are slightly different from each other. The estimated values are \( \omega_1 = 1.385 \) [Hz], \( \omega_2 = 1.376 \) [Hz], \( \omega_3 = 1.389 \) [Hz] and \( \omega_4 = 1.382 \) [Hz]. From the oscillation frequencies, the angular frequency is calculated by \( \omega_i = 2\pi f_i \). In addition, the values of parameter of the pendulums are \( \xi_1 = 0.0226, \xi_2 = 0.0228, \xi_3 = 0.0231, \) and \( \xi_4 = 0.0237 \). The parameter values of the board are \( \xi_p = 0.00113, \gamma = 0.024, \eta = 0.01\omega_p^2 / g, \) and \( \alpha = 1.072 \). In the numerical simulations, the absolute value of the angular velocity \( \omega \) of the pendulum is increased by 25.8 [deg/s] due to the impulsive force of the metronome when the angle \( \theta_i \) becomes \( \pm 10^\circ \). Then, metronome can continue to oscillate.

3. Results

3.1. Time-series data

Figures 2 and 3 show the time-series of the angle of each metronome. The horizontal axis is time [s], and the vertical axis is the angle \( \theta_i \) [deg].
In Fig. 2, the initial angles of the four metronomes are $\theta_1(0) = 60^\circ$, $\theta_2(0) = 20^\circ$, $\theta_3(0) = -20^\circ$, and $\theta_4(0) = -60^\circ$. Although the amplitudes of $\theta_2$ and $\theta_3$ are small at the initial state, their amplitudes reach to their maximum level which is about $60^\circ$ when the time is around $8 \text{ [s]}$. After $40$ second later, the four metronomes starts to synchronize. In Fig. 3, $\theta_1(0) = \theta_2(0) = \theta_3(0) = \theta_4(0) = 60^\circ$. Although the initial angles of all metronomes are set to $60^\circ$, they start to separate from each other when the time is around $8 \text{ [s]}$ because of the individual differences. We can then observe that a specific pair of metronomes synchronizes when $40 \leq t \leq 48$. However, their synchronized metronomes separate after $96$ seconds, and then a different pair of metronomes synchronize.

3.2. Relationship between frequencies of the metronomes and the board

![Figure 4: Relations between the mean frequency $f_s$ of the metronomes in the synchronous state and the frequency $f_p$ of the board.](image)

In Fig. 4, the circles indicate the in-phase synchronization, the triangle indicates the state that the in-phase synchronization and the anti-phase synchronization coexist, and the squares indicates the anti-phase synchronization. The horizontal axis shows the frequency of the board, and the vertical axis shows the average frequency of the four metronomes.

As shown in Fig. 4, when the frequencies of the metronomes are equal to that of board, $f_s$ reaches the largest value. From these results, when $f_s$ is large, the metronomes acquire energy from the board. When the frequencies of the metronomes approaches to that of the board, the metronomes are likely to acquire energy from the board.

3.3. Time necessary for the in-phase synchronization

We next investigate the time required for the in-phase synchronization when we change the frequency of the board. In these numerical experiments, we set the initial states of the metronomes to $\theta_1(0) = 60^\circ$, $\theta_2(0) = -60^\circ$, $\theta_3(0) \in [-60^\circ, 60^\circ]$, and $\theta_4(0) \in [10^\circ, 60^\circ]$.

![Figure 3: The same as Fig.1, but for $f_p = 1.600[\text{Hz}]$.](image)
Figure 5: Time necessary for the in-phase synchronization when metronomes are identical.

Figure 6: Time necessary for the in-phase synchronization when metronomes are different.
4. Conclusion

We fixed the frequencies of four metronomes to 1.4[Hz] in this study and performed numerical simulation of synchronization of four metronomes. When the frequency of the metronomes approaches to that of the board, the metronomes are likely to acquire energy from the board. When the frequency of the metronome is larger than that of the board, we showed that metronomes synchronize within approximately 40 seconds. In addition, we showed examples of the phase set to two specific initial states when the frequency of the metronome is smaller than the frequency of the board. We clarified that the frequency of the metronomes reaches the highest value when the natural frequency of the metronome is close to that of the board. In case of four metronomes, we can set initial angles into two pairs of symmetric angles in comparison with an experiment in the synchronization of three metronomes[5]. We confirmed that time necessary for the in-phase synchronization becomes shorter if there was no individual difference. Furthermore, we confirmed that the initial angles which accomplish the synchronization randomly exist when the frequency of the metronomes and the frequency of the board are identical. When the frequencies of the metronome and the board are identical and the metronomes synchronize, the frequency of the metronomes increases. However, the range of synchronization with no individual difference becomes small. Therefore, existence of the individual difference plays an important role for synchronization.

In this paper, we conducted only numerical simulations of the four metronomes. However, it is inevitable to compare numerical results and experimental results. As for the case of three metronomes, we have already reported several interesting results[5]. It is an important future issue to analyze the motion of four metronomes experimentally.

Acknowledgments

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. is supported by Grant-in-Aid for Research Activity Start-up (No. 26880020) and by Grant-in-Aid for Young Scientists B (No. 16K16126) from JSPS. The research of K.F. was supported by Grant-in-Aid for Challenging Exploratory Research (No. 15K12137) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Challenging Exploratory Research (No. 24650116) from JSPS and by Grant-in-Aid for Scientific Research (C) (Generative Research Fields) (No. 15KT0112) from JSPS.

References


Comparison of Resource Budget Models for Nonlinear Dynamics in Alternate Bearing of Tree Crops

Xujun Ye† and Kenshi Sakai‡

†Faculty of Agriculture and Life Science, Hirosaki University, 3 Bunkyo-cho, Hirosaki, Aomori 036-8561, Japan
‡ Environmental & Agricultural Engineering Department, Tokyo University of Agriculture and Technology, 3-5-8 Saiwai-cho, Fuchu, Tokyo 183-8509, Japan
Email: yexujun@hirosaki-u.ac.jp, ken@cc.tuat.ac.jp

Abstract— This paper presents a comparative study of three recent models, i.e. the original Isagi’s model [1] and two recent models by Ye and Sakai [2,3], for modeling the alternate bearing of tree crops. The theories, definitions and performance of these models are discussed. All of these models are based on theoretical assumptions of resource allocation and budgeting in plant body. Simulations indicate that the new models may be more suitable for modeling the alternate bearing phenomenon in evergreen tree crops such as citrus plants.

1. Introduction

Alternate bearing or masting is commonly found in tree crops. Scientists have proposed several hypotheses for explaining the phenomenon from the perspective of internal resource allocation and budgeting in plant. Isagi et al. (1997) first proposed the resource budget model and tested it with experimental data for F. crenata [1]. Recently, Ye and Sakai have developed two new models based on Isagi’s model by incorporating vegetative growth component and introducing new equations for describing the relationship between vegetative and reproductive growths [2,3]. This paper attempts to explain the theories and definitions of these models for a better understanding of the models for this phenomenon.

2. Model theories and definitions

The Isagi’s model [1] describes a deterministic process of energy storage due to photosynthesis, and energy depletion due to fruit and flower production in the plant. It hypothesizes that each crop accumulates photosynthesize every year, producing flowers and sets fruits at a rate limited by the ratio between costs for flowering and fruiting in individual plants.

Ye and Sakai (2016) assume that the growth of new leaves (\(C_t\)) in early season constitutes a major part of energy consumption, which is determined by a ratio (\(r\)) of the potential resource accumulation capacity that depends on the reproductive threshold (\(L_T\)), the energy reserve (\(I\)) and the new resources accumulated by photosynthesis (\(P_s\)) in the year:

\[
C_t(t) = r(L_T + P_s - I(t))
\]

Further, to add nonlinearity to the model, a Ricker-type equation is used to describe the relationship between flowering (\(C_f\)) and fruiting (\(C_a\)) events in the model:

\[
C_a(t) = R_a C_f(t) e^{\frac{a(C_f(t))}{K}}
\]

where \(a\) represents the fruiting coefficient; \(K\) represents the optimal cost of flowering.

In another new model [3], the inter-dependent phenological relationships between vegetative and reproductive growths are highlighted. The plant growth dynamics is described by two equations:

\[
C_f(t) = \frac{cP_f}{1 + e^{(-bc(t)-a)}}
\]

\[
C_f(t+1) = \frac{C_f(t)^2}{C_f(t) + d} + kC_a(t)
\]

where \(a, b, c, d, k\) are constants, and \(0<k<1\). Natural defoliation is determined by \(d\), and new leaf growth is \(k\) times preceding cost for fruiting. \(k\) represents the degree of enhancing effect of fruiting (\(C_a(t)\)) on subsequent new leaf growth (\(C_f(t+1)\)).

3. Conclusion

This paper presents a comparative study of three recent models developed for alternate bearing of tree crops. All of these models are based on theoretical assumptions of resource allocation and budgeting in plant body and can model the alternate bearing or masting in tree crops. Simulations indicate that the new models may be more suitable for evergreen tree crops such as citrus plants.

References

Understanding the some aspects of Alternate Bearing Phenomenon: cycle of three years

Awadhesh Prasad¹, Kenshi Sakai² and Yoshinobu Hoshino³

¹Department of Physics and Astrophysics, University of Delhi, Delhi 110007, India
²Department of Environmental and Agricultural Engineering, Tokyo University of Agriculture and Technology, Tokyo 183-85-9, Japan
³Department of Environment Conservation, University of Delhi, Delhi 110007, India

Email: awadhesh@physics.du.ac.in, ken@cc.tuat.ac.jp, hoshino@cc.tuat.ac.jp

Abstract– Alternate bearing is a common phenomenon in plants having big seed where fruiting occurs in alternate years i.e. cycle of two years. In this paper we explore the possibility of three years cycle instead of two years. Nonlinearity is used in energy resource model to understand this enhanced cycle.

1. Introduction

We have considered a well known/studied system which captures many of the alternate bearing phenomena. However other variants may be considered for generalization of the presented results. Experimenting this observed results are also possible. However it will take years/decades to check the usefulness of these results. However in this work a made an attempt and discuss the feasibility for experimentation in real field. Question is why does it happen? Is it possible to model such behavior using the energy resource [1] which is based on energy generated by photosynthesis? Answer to these questions are the main aim of this present work.

2. Model

A constant amount of photosynthate is produced every year in individual plant. This photosynthate is used for growth and maintenance of the plant. The remaining photosynthate (P) gets stored in the plant body. The accumulated photosynthate stored in plant is expressed as I. In a year when accumulated photosynthate exceed a certain threshold L₁ then the remaining amount, I = L₁, is used for flowering as the cost of flowering, C. These flowers are pollinated and bear fruits which cost is designated as Cᵢ. Usually, the fruiting cost is a function of the cost of flowering, i.e., Cᵢ = F(C) where F(.) is a function. The accumulated photosynthate becomes I = L₁ after flowering. Once fruiting is over it becomes L₁ = Cᵢ, i.e., I = L₁ = F(Cᵢ). Under the less fluctuating conditions we may consider the function F(Cᵢ) as a linear one i.e., Cᵢ = RCᵢ where R is a constant. In this case this phenomenon is modeled as [1] this is termed as Isagi Model or Resource Budget Model which is given as

\[ L_{int} = -RL₁ \]
\[ I₁ > L₁ - P₁ \]
\[ = I₁ + P₁ \]
\[ I₁ ≤ L₁ - P₁ \] (1)

where \( L = L₁ - (1+R)P₁ \). This model has been extensively studied in literature [1,2]. Let us consider the situation when function F(Cᵢ) is nonlinear, which varies with Cᵢ as Cᵢ ≈ Cᵢ^β i.e. Cᵢ = RCᵢ^β; where R is a proportionality constant and β is the scaling factor determining the nonlinearity. In the presence of this nonlinearity, the Model, Eq. (1), becomes

\[ I_{int} = L₁ - R(I₁ + P₁ - L₁)^β \]
\[ I₁ > L₁ - P₁ \]
\[ = I₁ + P₁ \]
\[ I₁ ≤ L₁ - P₁ \] (2)

3. Results and discussions

Results are shown in Fig. 1. The bifurcation diagram in Fig. (1) shows that if \( α < 1.15 \) the motion is periodic while afterward it is chaotic. An important observation is the occurrence of period-3 type window near \( α=1.4 \) which trajectory is shown in Fig. 1(b). It shows that once there is heavy fruiting then there is certain that there is no fruiting in the following year. Whoever there is intermediate fruiting before heavy year one. This happens as certain energy (photosynthate) after intermediate yield remains available for next year for heavy fruiting. This suggests that there is possibility of three years cycles instead of two years if nonlinear variation in Cᵢ is considered.

Fig.1. (a) Bifurcation diagram as a function R and (b) trajectory at R=1.4 for fixed β=0.5, L₁=1 and P₁=0.5.

Acknowledgments: AP acknowledges support from JSPS Invitation Fellowship(IF02) S15093. KS thanks to Dr. Tohru Ikeguchi with his valuable discussions and JSPS Grants-in-Aid 23380152 and 15KT0112.
References


Spatial Dynamics of Acorn Masting and Tree Crops Alternate Bearing

Kenshi Sakai1 and Yoshinobu Hoshino2

1Department of Environmental and Agricultural Engineering, 2Department of Environment Conservation, Faculty of Agriculture, Tokyo University of Agriculture and Technology, Tokyo 183-85-9, Japan

Email: ken@cc.tuat.ac.jp, hoshino@cc.tuat.ac.jp

Abstract

In this paper, we investigate spatial synchronization of acorn masting and alternate bearing. Spatially extended resource budget models are used to demonstrate spatial dynamics of masting and alternate bearing. Pollination is considered as a coupling term of the model. Global coupling and local coupling are introduced by mean field and coupled map lattice, respectively. We also review the related researches in terms of field applications.

1. Introduction

Acorn masting is natural phenomenon by which plant populations occasionally produce a large number of seeds. It has been an attractive subject in wide areas [1,2]. It is one of subjects to investigate in wild life management [3], as acorns are substantial food for wild animals. In sylviculture, to predict the tendency of masting in forest stands is a key technology for promoting natural regeneration approach. Acorn masting is recognized as a synchrony of collective dynamics consisting of coupled nonlinear oscillators and many types of coupling are investigated [4]-[8]. Alternate bearing of tree crops has been central subject in agriculture and pomology. Citrus (e.g. Orange, Lemon, Mandarin), Pistachio, Chestnut are the crops show pronounced alternate bearing[9]-[13].

2. Model

Isagi’s resource budget model (RBM) can be used to explain the masting in whole forest stands and alternate bearing of each tree[4]. Combining RBM and coupled map lattice, the spatial dynamics of acorn masting was demonstrated e.g.[14]. The effect of weather conditions known as Moran’s effect was investigated in terms of common noise induced synchrony[10][15][16]. Many attempts of modification of RBM for real world applications has been done e.g.[17][18].

3. Sensing

Sensing technologies of acorn and crop yield in forest stands and orchards are expected not only for field application but also validation of proposed masting models. Remote sensing techniques have been introduced in this purpose. Multi-spectral imaging and hyperspectral imaging were employed in airborne remote sensing [19][20] and ground truth[21].

4. Measures

Spatial autocorrelation is one of popular measures for quantifying the spatial synchronization. It was applied on acorn [14] and pistachio [11]. In terms of nonlinear dynamics, nonlinear time series analysis tools are also expected to evaluate the spatial dynamics. For example, Lyapunov exponents(LE), translation error(TE), deterministic nonlinear prediction(DNP) might be potential candidates to be modified for it.

5. Implementation

For field applications, GIS embedded models[15] is expected to demonstrate real world masting and alternate bearing. Both of canopy identification and calibration methods are the most important options to be developed.

6. Conclusions

At the present stage, the skeleton such as models, sensing technologies and prediction(DNP) / control (OGY)[22] has been mostly prepared for field applications in masting and alternate bearing. It is necessary to set up collaborative works in inter-disciplinary / inter-sectional ways for realizing successful applications.

References

[16] Sviridova & Sakai, JJSAM, 78(3), 221, 2016
[22] Sakai & Noguchi, Chaos Solitons Fractals, 41(2), 630, 2009
Fredholm Determinants of Generalized $\beta$-Transformations and MSE Estimates of Corresponding AD-Converters

Katsutoshi Shinohara$^\dagger$ and Kenta Koayashi$^\ddagger$

$^\dagger$Faculty of Commerce and Managements, Hitotsubashi University
2–1 Naka, Kunitachi, Tokyo 186-8601, Japan
Email: ka.shinohara@r.hit-u.ac.jp, kenta.k@r.hit-u.ac.jp

Abstract—We consider an Analog-to-Digital (AD) encoder defined by generalized $\beta$-transformation. Such encoders are called $\beta$-encoders. In this article, we show that the mean squared error of a $\beta$-encoder can be estimated by analyzing the zeros of the Fredholm determinant of the transformation. We give an example of a rigorous upper bound of the MSE by this technique, together with the numerical verification method.

1. Introduction

A $\beta$-encoder is an Analog-to-Digital (AD) encoder based on the $\beta$-expansion. Compared to the conventional encoders based on the binary expansion, it is proposed that they show self-correcting property ([1]) and accordingly are advantageous for designing less energy consumption and smaller encoders. Experiments witness these advantages (see for instance [2]). While $\beta$-encoders have nice properties from the viewpoint of engineering, their mathematical treatment contains a lot of challenging problems.

In this paper, we consider the problem of finding the upper bound of mean squared error (MSE) of such AD-converters. An analog input is encoded to the digital output through $\beta$-encoders. In practice, we can only maintain finite number of digits thus there is loss of information. For optimizing the design of the encoder, it is important to establish better upper bound of MSE.

This problem is translated to estimate a certain integral of a function given by the iteration of transfer curve of the AD-encoder under consideration. This is an easy problem for conventional AD-converters since the points of discontinuity distribute evenly over the interval. On the other hand, this is a very complicated problem for $\beta$-encoders: In general, the distribution of the points of discontinuity are scattered over the interval unevenly. Furthermore, since the transfer curve has positive Lyapunov exponent $\log \beta$, the iteration exhibits the sensitive dependence on the initial condition as the number of iteration increases, that is, when we consider $\beta$-converter with large number of digits.

As a result, the behavior of the MSE shows quite a complicated behavior when the parameter value $\beta$ varies.

A simple MSE upper bound for $\beta$-encoders is obtained in [3], by means of change of the variable argument and approximation of integrals by a Markov process of finite states. The result is rigorous. However, since the proof is done by approximation method, as the number of digits increases the amount of calculation required for obtaining the inequality increases. Consequently, the conclusion we could obtain was limited. We gave MSE only for the case for limited number of digits.

In this article, we propose a new method for estimating the MSE of $\beta$-encoders. It is based on the analysis of the eigenvalues of the Perron-Frobenius operator of the transformation. Roughly speaking, in this method we consider the statistic property of infinite-states Markov process directly, that is, without approximation. The advantage of this method is that the size of calculation does not change as the number of digits increases. Indeed, by this method we can derive an upper bound of MSE valid for all (sufficiently large) digits.

To establish the inequality, we need to estimate two quantities. Namely, the absolute value of the second eigenvalue of Perron-Frobenius operator and the coefficient of the corresponding decay term. By numerical verification method, we can give a rigorous upper bound for these quantities. Because of the limitation of pages, we do not discuss much about this part in this paper.

The organization of this article is as follows: In section 2 we present basic definitions and give the precise statement of our problem. In section 3 we review the calculation of Fredholm determinant of the Perron-Frobenius operator of $\beta$-transformations. In section 4 using the result of section 3 we give an upper bound of MSE. In section 5 we present the numerical result.

2. Precise statement

2.1. Setting

Let $\beta \in (1, 2]$ and $\nu \in [1 - \beta^{-1}, \beta^{-1}]$. We consider the transformation on $I = [0, 1]$ defined as follows (notice that it is slightly different from the usual $\beta$-transformation):

$$T_\beta,\nu(x) = T(x) = \begin{cases} \beta x & (x \leq \nu) \\ \beta(x - 1) + 1 & (x > \nu). \end{cases}$$

For $x \in I$, we also define the $i$-th digit $d_i(x)$ of $x$ by

$$d_i(x) = \begin{cases} 0 & (T^{i-1}x \leq \nu) \\ 1 & (T^{i-1}x > \nu). \end{cases}$$
Then, the infinite sequence \((d_i(x)) \in [0, 1]^\mathbb{N}\) gives a \(\beta\)-expansion of \(x\), namely, we have the following equality:

\[
x = (\beta - 1)^{-1} \left( \sum_{i=1}^{\infty} d_i(x) \beta^{-i} \right),
\]

(1)

The coefficient \((\beta - 1)^{-1}\) is the normalizing constant. The sequence \((d_i(x))\) is the digital encoding of an input \(x\) and by (Eq. 1) we can recover the input from the infinite \([0, 1]\)-sequence.

In an ideal situation where all the (infinitely many) digits are available, the encoding and decoding process does not bring any loss of information. However, in the real world only finitely many digits are available. By \(L \in \mathbb{N}\) we denote the number of digits available. Then for an input \(x\), to recover its original value from the quantization \((d_i(x))\), instead of (Eq. 1) we can recover the input from the infinite sequence.

For fixed \(x\) only finitely many digits are available. By \(L\) we denote any loss of information. However, in the real world only finitely many digits are available. By \(L \in \mathbb{N}\) we denote the number of digits available. Then for an input \(x\), to recover its original value from the quantization \((d_i(x))\), instead of (Eq. 1) we can recover the input from the infinite sequence.

Then, by a change of variable argument, we have the following equality:

\[
\text{MSE}(\beta, \nu, L) = \beta^{-2L} \sum_{i=0}^{2L-1} n_i^{(L)} I_i,
\]

where \((I_i)\) are real numbers given as follows: We put \(J_0 = [\nu, r_1]\). Then, \(I_i = \frac{1}{2} (r_i - \theta^i) - (r_i - \theta^{i+1})\). Thus if we have some estimate about the behavior of \(n_i^{(L)}\) then we obtain an upper bound of MSE.

Intuitively, we may guess that for fixed \(\beta\) and \(\nu\), MSE has order \(\beta^{-2L}\). Thus in the following we are interested in estimating the following constant:

\[
K_{\beta, L} := \sum_{i=0}^{2L-1} n_i^{(L)} I_i,
\]

(3)

3. Perron-Frobenius operator and segments

3.1. Perron-Frobenius operator

For a piecewise \(C^1\) transformation \(S\) of an interval \(I\), we can define the Perron-Frobenius operator \(P\) acting on \(L^1(I)\) as the (extension of) the adjoint operator of the Koopman operator \(f(x) \mapsto f(S(x))\) with respect to the \(L^2\) inner product. We are interested in calculating \(P(\mathbb{1}_J)\) where \(P\) is the Perron-Frobenius operator for \(T\) introduced in the previous subsection and \(\mathbb{1}_J\) is the characteristic function of an interval \(J \subset I\). For such functions, we have

\[
P(\mathbb{1}_J) = (1/\beta) (\mathbb{1}_{T(J_-)} + \mathbb{1}_{T(J_+)})
\]

where \(J_- = J \cap [0, \nu]\) and \(J_+ = J \cap [\nu, 1]\).

Notice that by definition we have

\[
P^L(\mathbb{1}_J) = \beta^{-L} \sum_{i=0}^{2L-1} n_i^{(L)} \mathbb{1}_J.
\]

Thus, the analysis of the MSE is reduced to the study of corresponding Perron-Frobenius operator. In the following, we investigate the behavior of the sequence of functions \((P^L(\mathbb{1}_J))\).

3.2. Generating function

To analyze \((P^L(\mathbb{1}_J))\), we introduce a nice tool to analyze sequences satisfying recursive relations called generating function. Consider the following formal power series of functions

\[
s'(z) := \sum_{i=0}^{\infty} (P(\mathbb{1}_J)) z^i,
\]

where \(z\) is a formal variable. Notice that formally we have \(s'(z) = (1 - zP)^{-1}(\mathbb{1}_J)\). This suggests that the solution of the equation \(s'(z) = 0\) is the reciprocal of the eigenvalue of the operator \(P\).
3.3. Fredholm determinant

Using the recursive relation of the sequence \((P^i(1))\), we can derive a closed formula of \(s^i(z)\). This is done for the greedy \(\beta\)-transformation by Ito and Takahashi [4]. For general case, this is done by Mori [5]. For simplicity, in the following we assume that \(\nu = \beta^{-1}\), that is, the left branch of \(T\) covers the whole \(I\). In this case, in each iteration of \(T\) there is at most only one non-empty segment. Thus we forget the a priori empty segments and denote the non-empty ones as \(J_0, \ldots, J_L\). We choose \(\theta\) to be the middle point of \(J_1\), that is, \(\theta = (3 - \beta)/2\). In this setting, \(s^i(z)\) is given by the following formula:

\[
s^i(z) = \frac{1}{(1 - z/\beta)(1 - E(z))} \left( \sum_{s=1}^{\infty} (z/\beta)^s L_s \right),
\]

where

\[
E(z) = \sum_{i=1}^{\infty} \phi(i - 1) \left( \frac{z}{\beta} \right)^i, \quad \phi(i) = \begin{cases} 0 & (T^{r+1}(1/\beta) \geq 1/\beta), \\ 1 & (\text{otherwise}). \end{cases}
\]

Notice that, while the domain of convergence of \(s^i(z)\) was initially \(|z| < 1\), \(s^i(z)\) converges for \(|z| < \beta\) in the new formula. Thus we have obtained an analytic continuation of \(s^i(z)\).

By taking the Taylor expansion of \(s^i(z)\), we can extract some information about \(n_i^{(L)}\). We put

\[
\frac{1}{(1 - z/\beta)(1 - E(z))} = \sum_{i=0}^{\infty} w_i z^i.
\]

By expanding (Eq:4) and comparing the coefficients, we obtain

\[
n_i^{(L)} = w_{L-i} \cdot \beta^{L-i}.
\]

Thus, in order to obtain the estimate of MSE, we need to know the behavior of the sequence \((w_i)\).

3.4. Taylor expansion of coefficient function and MSE

Let us estimate the coefficients \((w_i)\). In our setting, we can prove that \(z = 1\) is a simple root of \(1 - E(z) = 0\) (see Mori for example).

Thus we have the following factorization:

\[
1 - E(z) = (1 - z)R(z),
\]

where \(R(z)\) is a holomorphic function on \(|z| < \beta\). By a simple calculation together with the fact that \((\phi(i))\) is related to the \(\beta\)-expansion of \(1/\beta\), we have

\[
R(z) = \sum_{i=0}^{\infty} \frac{1}{(\beta - 1) \beta^i} (T^{r+1}(1/\beta)^i z^i).
\]

Thus, (Eq:5) is equal to

\[
\frac{1}{(1 - z/\beta)} \left[ \frac{1 - rR(z)}{(1 - z)R(z)} \right] + \frac{r}{(1 - z/\beta)(1 - z)},
\]

where \(r = 1/R(1)\).

The Taylor expansion of these terms provides us with information of \((w_i)\). For instance, the coefficient of \(z^n\) from the second term (we denote it by \(u_n\)) is

\[
u_n = \frac{r}{\beta - 1} (\beta - \beta^n).
\]

Let us estimate the coefficient of the first term.

3.5. Second eigenvalue and estimate of coefficient

In order to obtain the estimate of contribution of the first term, we use contour integrals. We denote the coefficient of \(z^n\) of the first function by \(v_n\). Notice that the function in the integral has no pole in \(|z| \leq \mu\) for every \(\mu < \eta\), where \(\eta\) is the absolute value of zero of \(R(z) = 0\) with smallest absolute value. Thus we have the following equality.

\[
v_n = \frac{1}{2\pi i} \int_{|z| = \mu n \cdot \eta} \frac{1 - rR(z)}{(1 - z)(1 - z/\beta)R(z) \cdot z^{n+1}} d\theta.
\]

Thus if \(\mu\) can be chosen greater than 1, then we know that \(v_n\) decays exponentially:

\[
v_n \leq \frac{J(\mu)}{2\pi} \mu^{-(n+1)},
\]

where

\[
J(\mu) = \left| \int_{|z| = \mu n \cdot \eta} \frac{1 - rR(z)}{(1 - z)(1 - z/\beta)R(z)} d\theta \right|.
\]

These constants can be calculated numerically. Indeed, adopting numerical verification method, we can establish an upper bound for \(\eta\) and once we fix \(\mu\), then it is possible to obtain the upper bound of \(J(\mu)\) (as an upper bound of the contour integral). Together with these constants with rigorous numerical verification, we can derive several rigorous upper bounds of MSEs.

4. Estimation of MSE

4.1. Calculation of population

Recall that, in order to obtain the upper bound, we only need to obtain the upper bound of the constant \(K_{\beta,L}\) in (Eq:3). Using (Eq:6) and \(w_n = u_n + v_n\), we have

\[
K_{\beta,L} = \sum_{i=0}^{L} \frac{r}{\beta - 1} (\beta - \beta^{-(L-i)}) \frac{I_i}{\beta^i} + \sum_{i=0}^{L} v_{L-i} \frac{I_i}{\beta^i}.
\]

Let us consider the case where \(L\) is sufficiently large (say \(L \geq 18\)), since the case where \(L\) is smaller than these values are treated in the paper [3].

To obtain the estimate of \(K_{\beta,L}\), we divide the sum into two parts: the part \(i \leq 10\) and the part \(i > 10\). We denote the former one by \(K_{10}\) and the other by \(K_{\infty}\). The term \(K_{10}\) is not hard to estimate since we consider \(\beta\) in a small interval. Hence we calculate it directly.
The latter part is hard to estimate since it is related to the dynamics of higher iteration. Thus we only give an upper bound to it. The quantity we want to estimate is:

\[
\sum_{i=1}^{L} \frac{r_i I_i}{\beta - 1} = \sum_{i=1}^{L} \frac{r_i I_i}{\beta - 1} B^i + \sum_{i=1}^{L \rightarrow L_i} \frac{I_i}{\beta} B^i.
\]

For the second term, since it is negative and small we neglect it. For the first and the third term, we substitute \( I_i \) with its worst value \( \tilde{I} \), that is, we assume \( I_i = 0 \) and \( r_i = 1 \). Namely, recalling that we put \( \theta = (3 - \beta)/2 \),

\[
\tilde{I} = \frac{1}{3} \left[ (1 - \frac{1}{2}(3 - \beta))^3 - (\frac{1}{2}(3 - \beta))^3 \right] = \frac{1}{3}((2\beta^2 - 12\beta + 13).
\]

By this substitution, for the first term we have

\[
\sum_{i=1}^{L} \frac{r_i}{\beta - 1} B^i \leq \frac{r}{\beta^1} B^1 = \frac{r \tilde{I}}{\beta^1}.
\]

For the third term, we use the upper bound obtained by the second eigenvalue. By (Eq:7), for every \( 11 \leq i \leq L \) we have

\[
|v_{L-i} I_i| \leq \frac{J(\mu)}{2\pi \mu^{L-i+1}} \tilde{I} / \beta \leq \frac{J(\mu) \tilde{I}}{2\pi \mu} \left( \frac{\mu}{\beta} \right)^L =: K_\mu.
\]

Accordingly, we have

\[
\left| \tau_{L-i} \right| \frac{I_i}{\beta} \leq \frac{J(\mu) \tilde{I}}{2\pi \mu^{L+1}} \left( \frac{\mu}{\beta} \right) \sum_{i=1}^{L} \left( \frac{\mu}{\beta} \right)^i
\]

\[
\leq \frac{J(\mu) \tilde{I}}{2\pi \mu^{L+1}} \left( \frac{\mu}{\beta} \right) \sum_{i=1}^{\infty} \left( \frac{\mu}{\beta} \right)^i
\]

\[
\leq \frac{J(\mu) \tilde{I}}{2\pi \mu^{L+1}} \left( \frac{\mu}{\beta} \right)^{\infty} = \frac{J(\mu) \tilde{I}}{2\pi \mu^{L+1}} \left( \frac{\mu}{\beta} \right)^L =: K_\mu.
\]

Finally, we have \( K_{\beta,L} = K_{10} + K_{\alpha} \leq K_{10} + K_\mu + K_b \). Notice that, compared to \( K_{10} \) the other two terms are very small. Thus \( K_{10} \) is the dominant term of \( K_{\beta,L} \). We also remark that, by examining the calculation carefully, the above upper bound is valid as the upper bound for \( K_{\beta,M} \) for every \( M \geq L \). Thus our result provides the upper bound not only for specific \( L \) but also for every \( K_{\beta,M} \) with \( M \geq L \).

5. Numerical result

5.1. Second eigenvalues

For the sake of simplicity, we concentrate on the case where \( \beta \in [1.83, 1.8300001] \). We believe that the same technique would provide similar results with more calculation. In this case, we can prove the following:

- For the \( \beta \) above, let \( \beta \) be the root of \( R(z) = 0 \) with the smallest eigenvalue. Then we have \( |\beta| > 1.622531 \).
- Letting \( \mu = 1.55 \), we have \( J(\mu) < 4.774380 \).

5.2. The estimation of MSEs

By these results, combining the result in the previous section, we obtain the following:

- For the \( \beta \) above, we have \( 0.0449 < K_{10} + K_{\alpha} < 0.0450 \).
- The error between this value and numerically estimated \( K_{\beta,20} \) is less than 1%.

6. Summary

By means of spectral analysis of Perron-Frobenius operators, we derived an upper bound for MSE of \( \beta \)-encoders. This upper bound is valid not only for some specific number of digits but also for every sufficiently large digits. We believe that by this method we can establish the upper bound not only for \( \beta \) in a narrow range considered in this paper but also for every digit in the other parameter range or for encoders given by different threshold. We would like to complete such research in the other opportunities.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 16K00333.

References


Pseudo Random numbers generated by Dynamical Systems

Makoto Mori

Department of Mathematics, College of Humanities Sciences, Nihon University
3–25–40 Sakura Jyosui, Setagayaku, Tokyo 156-8550, Japan
Email: mori@math.chs.nihon-u.ac.jp

Abstract—One of the most important pseudo random numbers is a van der Corput sequence. We will consider it from the view point of dynamical systems, and show that the discrepancy of pseudo random numbers is deeply connected with the ergodicity of a dynamical system. We first study 1–dimensional cases, and then construct higher dimensional transformations which generate low discrepancy sequences. Main tools are the spectra of Perron–Frobenius operator and renewal equations.

1. Introduction

Let \( I = [0, 1]^d \) (\( d \geq 1 \)) and \( F: I \to I \). We consider a partition \( \{ (a) : a \in \mathcal{A} \} \) of \( I \), and express \((I, F)\) to a symbolic dynamics. Let \( X \) be a set of infinite sequences of symbols \( a_1a_2\cdots (a_i \in \mathcal{A}) \), such that \( \cap_{n=1}^{\infty} F^{-n+1}(\{(a_i)\}) \) consists of unique point. We assume a dynamical system \( X \) with the shift is isomorphic to \((I, F)\).

We denote a finite sequence of symbols \( a_1\cdots a_n (a_i \in \mathcal{A}) \) a word and

- \( |w| = n \),
- \( \langle w \rangle = \bigcap_{i=1}^{n} F^{-i+1}(\{a_i\}) \),
- for \( x \in I \), \( wx \) is a point such that \( wx \in \langle w \rangle \) and \( F^{|w|}(wx) \equiv x \), if it exists.

We consider a some order on \( \mathcal{A} \), and define an order \( w \prec w' \) \((w = a_1\cdots a_n, w' = b_1\cdots b_n)\) if

- \( |w| < |w'| \),
- \( |w| = |w'| \), and there exists \( k \) such that \( a_{k+1}\cdots a_n = b_{k+1}\cdots b_n \) and \( a_k < b_k \).

We call a set \( [wx] \) with the above order a van der Corput sequence generated by the dynamical system \((I, F)\). The famous van der Corput sequence for binary case:

\[ 0.1, 0.01, 0.11, 0, 001, 0.101, 0.011, 0.111, \ldots \]

corresponds to \( d = 1 \), \( F(x) = 2x \) (mod 1) and \( x = \frac{1}{2} \).

2. pseudo random numbers

A sequence \( x_1, x_2, \ldots \in I \) is called uniformly distributed if

\[
D(N) = \sup_{J} \frac{1}{N} |\{ n \leq N : x_n \in J \} - |J||
\]

converges to 0 as \( N \to \infty \), where supremum is taken over all the intervals in \( I \) and \( |J| \) is the Lebesgue measure of \( J \).

It is conjectured that any sequence satisfies

\[
D(N) \geq O \left( \frac{(\log N)^d}{N} \right).
\]

Thus the sequences which satisfies the equality in the above inequality is called of low discrepancy, this means the low discrepancy sequences are the best possible pseudo random numbers. Ninomiya ([9, 10]) showed that the van der Corput sequences generated by \( \beta \)-transformation are of low discrepancy. In 1–dimensional cases, we will extend this result to more general piecewise linear cases, and at the same time, we will construct low discrepancy sequences in higher dimensional cases.

3. Perron–Frobenius Operator

We have constructed pseudo random numbers using dynamical system \((I, F)\). The discrepancy of these sequences are deeply connected with the spectra of the Perron–Frobenius operator \( P \) associated with the dynamical system:

\[
Pf(x) = \sum_{y:F(y)=x} f(y) |J(F)(x)|^{-1},
\]

where \( J(F)(x) \) is the Jacobian of \( F \) at \( x \). In the following we assume that \( |J(F)| \equiv \beta \) and \( \beta > 1 \). In terms of the symbolic dynamics, we can express \( P \) by

\[
Pf(x) = \sum_{a \in \mathcal{A}} f(ax) \beta^{-1}.
\]

The spectra of the Perron–Frobenius operator determine the ergodic properties of the dynamical system:

- \( 1 \) is the eigenvalue of the Perron–Frobenius operator, and we can choose a base of the eigenspace by density functions of the invariant measures of the dynamical systems.
- Assume that \( 1 \) is the simple eigenvalue, that is, there exists unique invariant probability measure \( \mu \). If there exists no eigenvalue modulus 1 except 1, then the dynamical system is mixing.

When we restrict the domain of \( P \) to a suitable space (in 1–dimensional cases, \( BV \), the set of functions with bounded
variations), as the first greatest eigenvalue determines the invariant measure, the second greatest eigenvalue in modulus determines the speed of convergence to equilibrium:
\[ \int f(x) g(F^n(x)) \, d\mu \rightarrow \int f(x) \, d\mu \int g(x) \, d\mu. \]

4. Renewal equation

In this section, we consider the case of 1-dimensional dynamical systems \((d = 1)\), and the partition \(\{(a)\}_{a \in \mathcal{A}}\) is a partition of \(I = [0,1]\) by intervals. We call \(\inf(a)\) and \(\sup(a)\) endpoints of \(\{(a)\}_{a \in \mathcal{A}}\). A point \(\inf(a)\) is called Markov if \(\lim_{n \to \infty} F(x)\) also belongs to the set of endpoints, and a point \(\sup(a)\) is called Markov if \(\lim_{n \to \infty} F(x)\) also belongs to the set of endpoints. If there exists a partition such that all the endpoints are Markov, then we call \(F\) Markov.

**Theorem 1** Assume that there exists no eigenvalues \(|z| > \beta^{-1}\) except 1. Let \(k\) be the number of non-Markov endpoints, then
\[ D(N) = O \left( \frac{\log N}{N} \right). \]
This says the pseudo random numbers can be of low discrepancy only if \(F\) is Markov.

5. Renewal Equations

To prove the above theorem, we use renewal equations. To show the outline of the proof, we consider the case of the \(\beta\)-transformation \(F(x) = \beta x \mod 1\), and \(\beta\) equals the golden number \(\frac{1 + \sqrt{5}}{2}\).

Let \(\mathcal{A} = [a, b]\) and \((a) = [0, \beta^{-1})\), \((b) = [\beta^{-1}, 1]\). We define for \(c\) either \(a\) or \(b\)
\[ s^{(c)}(z, x) = \sum_{n=0}^{\infty} \beta^n P^n 1_{(c)}(x) \]
\[ = (I - zP)^{-1} 1_{(c)}(x). \]
This suggest that the singularities of \(s^{(c)}(z, x)\) equal the reciprocals of the eigenvalues of \(P\).

Now we construct a renewal equation. Note that \(F(a) = I\). Then
\[ s^{(a)}(z, x) = 1_{(a)}(x) + \sum_{n=1}^{\infty} \beta^n P^{n-1} (P1_{(a)}(x)) \]
\[ = 1_{(a)}(x) + \sum_{n=1}^{\infty} \beta^n P^{n-1} \left( \sum_{y \in F^n(x)} 1_{(y)} \beta^{-1} \right)(x) \]
\[ = 1_{(a)}(x) + \beta^{-1} \sum_{n=0}^{\infty} \beta^n P^n 1_{(x)} \]
\[ = 1_{(a)}(x) + \beta^{-1} \sum_{n=0}^{\infty} \beta^n P^n (1_{(a)}(x) + 1_{(b)}(x))(x) \]
\[ = 1_{(a)}(x) + \beta^{-1} (s^{(a)}(z, x) + s^{(b)}(z, x)). \]

On the other hand, as \(F(b) = \langle a \rangle\), we get
\[ s^{(b)}(z, x) = 1_{(b)}(x) + \beta^{-1} s^{(a)}(z, x). \]
Thus we can express them into the following form:
\[ \begin{pmatrix} s^{(a)}(z, x) \\ s^{(b)}(z, x) \end{pmatrix} = \begin{pmatrix} 1_{(a)}(x) \\ 1_{(b)}(x) \end{pmatrix} + \begin{pmatrix} \beta^{-1} & \beta^{-1} \\ 0 & \beta^{-1} \end{pmatrix} \begin{pmatrix} s^{(a)}(z, x) \\ s^{(b)}(z, x) \end{pmatrix}. \]
This is a renewal equation for one of the simplest cases. We denote \(\Phi(z) = \begin{pmatrix} \beta^{-1} & \beta^{-1} \\ \beta^{-1} & 0 \end{pmatrix}\), and call it the Fredholm matrix. We get
\[ \begin{pmatrix} s^{(a)}(z, x) \\ s^{(b)}(z, x) \end{pmatrix} = (I - \Phi(z))^{-1} \begin{pmatrix} 1_{(a)}(x) \\ 1_{(b)}(x) \end{pmatrix}. \]
This suggests that the solutions of \(\det(I - \Phi(z)) = 0\) are the reciprocals of the eigenvalues of the Perron–Frobenius operator \(P\). Moreover, we can prove \(\det(I - \Phi(z))\) equals the dynamical zeta function
\[ \zeta(z) = \exp \left[ \sum_{n=1}^{\infty} \frac{z^n}{n} \sum_{p, F^n(p) = p} |F^n(p)|^{-1} \right]. \]
Thus this suggests that the singularities of the dynamical zeta function are also the reciprocals of the eigenvalues of the Perron–Frobenius operator \(P\).

Actually, we can prove the above conjectures are true when we restrict \(P\) to \(BV\). We can extend the results to non-Markov transformations. See for detail [3, 4]. Moreover, we can extend the results to higher dimensional cases[7].

6. Discrepancies

We apply the results of the former section to calculate the discrepancy of the van der Corput sequences. For an interval \(J \subset I\), then
\[ \# \{wx \in J : |w| = n\} = \sum_{\|w\|=n} 1_J(wx) = \beta^n P^n 1_J(x). \]
Thus for a word \(u\) such that \(F^{|u|}(x) = I\)
\[ \sum_{n=0}^{\infty} z^n \# \{wx \in \langle u \rangle : |w| = n\} = \sum_{|w|=n} z^n \sum_{\|w\|=n} 1_{(w)}(wx) \]
\[ = \sum_{n=0}^{\infty} z^n \beta^n P^n 1_{(u)}(x) = \sum_{n=0}^{\infty} z^n 1_{F^n(u)(x)} + z^{|u|} s^u(g, z, x). \]
On the other hand, if \(N\) equals the number of words for which the length is less than or equal to \(n\), then
\[ \frac{1}{N} \sum_{k=0}^{n} \# \{wx \in J : |w| = k\} = \frac{1}{N} \sum_{k=0}^{n} \beta^k P^k 1_J(x). \]
Thus we can calculate it using the \(k\)-th coefficient of \(s^u(g, z, x)\) \((k \leq n)\), and we get the proof of Theorem 1. See for detail [5, 6], and the computer simulation of this pseudo random numbers, see [1].
7. Higher Dimensional cases

We can also construct a renewal equation for \( d \geq 2 \). However, the essential spectral radius of the Perron–
Frobenius operator is usually greater than \( \beta^{-1} \). Thus it is very difficult to construct pseudo random numbers of low
discrepancy. We will construct it using irreducible polynomials.

We consider a \( d \)-dimensional irreducible polynomial 
\( p(\beta) \) on \( \mathbb{F}_2 \). We express by \( \mathcal{A} = [1, \beta, \ldots, \beta^{d-1}] \) the additive
group generated by \( [1, \beta, \ldots, \beta^{d-1}] \).

We identify \( \beta^k \) (\( 0 \leq k \leq d-1 \)) as \( \left( \begin{array}{c} a_0^k \\ \vdots \\ a_{d-1}^k \\ \end{array} \right) \in \mathcal{A} \) such that
\( a_d^k = 1 \) and \( a_j^k = 0 \) (\( i \neq k \)). Thus for \( x \in [0,1]^d \) with its
binary expansion \( 0.a_1 a_2 \cdots \) \( (a_k \in \mathcal{A}) \), we can identify it as a sequence of \( \mathcal{A} \). Instead of constructing \( F : [0, 1]^d \to [0,1]^d \), we will construct \( F : \mathcal{A}^d \to \mathcal{A}^d \).

Let \( A_i = \left( \begin{array}{c} 1 \\ \beta^{2i-1} \\ \beta^{2i} \\ \vdots \\ \beta^{d(i-1)/2-1} \end{array} \right) \) (\( 1 \leq i \leq d \)). and consider a ma-
trix \( (A_1, A_2, \ldots, A_d) \). Note that this matrix has inverse on
\( \mathbb{F}_2 \). We denote its inverse matrix by \( \left( \begin{array}{c} X_1 \\ \vdots \\ X_d \end{array} \right) \), where \( X_i \) is a
d\(-\)-dimensional row vector.

**Example 1** For \( d = 3 \) and \( p(\beta) = 1 + \beta + \beta^3 \),
\[
(A_1, A_2, A_3) = \left( \begin{array}{ccc} 1 & 1 & 1 \\ \beta & \beta^2 & \beta^4 \\ \beta^2 & \beta^5 & \beta^8 \end{array} \right)
= \left( \begin{array}{ccc} 1 & 1 & 1 \\ \beta & \beta^2 & \beta + \beta^2 \\ \beta^2 & \beta^5 & \beta^2 + \beta^2 \end{array} \right),
\]
and
\[
\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \left( \begin{array}{ccc} 1 & \beta^2 & \beta \\ \beta & \beta^2 & \beta^4 \\ \beta^2 & \beta^4 & \beta^8 \end{array} \right)
\begin{bmatrix} \beta^2 & \beta \\ \beta^5 & \beta^2 \\ \beta^2 & \beta + \beta^2 \end{bmatrix}.
\]

We will define infinite dimensional matrices \( U = (a_{ij})_{i,j \geq 1} \)
and \( V = (x_{ij})_{i,j \geq 1} \), where \( a_{ij} \) is a \( d \)-dimensional column
vector of \( \mathcal{A} \) and \( x_{ij} \) is a \( d \)-dimensional row vector of \( \mathcal{A} \).
Note that in \( U \), 0 means the \( d \)-dimensional zero column
vector, and in \( V \), 0 means the \( d \)-dimensional zero row vec-
tor. Let us define rule A by
\[
\tilde{a}_{ij} = \begin{cases} \tilde{a}_{i-1,j-1} & \text{if } i,j = 1 \pmod{d}, \\
\tilde{a}_{i-1,j-1} + \tilde{a}_{i,j-1} & \text{otherwise}, \end{cases}
\]
with initial condition \( \tilde{a}_{11} = 1 \) and \( \tilde{a}(0,i) = 0 \), and we define a
matrix \( U \) by
\[
a_{ij} = \begin{cases} A_k & \text{if } \tilde{a}_{ij} = 1 \text{ and } j = k \pmod{d}, \\
0 & \text{if } \tilde{a}_{ij} = 0. \end{cases}
\]

Let \( \tilde{x}_{ij} \) also satisfy rule A with initial condition \( \tilde{x}_{ij} = 1 \) if \( \lfloor \frac{i}{d} \rfloor + j = 0 \pmod{d} \), and \( \tilde{x}_{ij} = 0 \) if \( \lfloor \frac{i}{d} \rfloor + j < j \). We define a matrix \( V \) by
\[
x_{ij} = \begin{cases} X_k & \text{if } \tilde{x}_{ij} = 1 \text{ and } i = k \pmod{d}, \\
0 & \text{if } \tilde{x}_{ij} = 0. \end{cases}
\]

**Example 2** For \( d = 3 \), \( U \) equals
\[
U = \begin{bmatrix} A_1 & A_2 & A_3 \\ 0 & A_2 & A_3 \\ 0 & 0 & A_2 \\ 0 & 0 & 0 \\
0 & 0 & 0 & A_2 \end{bmatrix},
\]
and \( V \) equals
\[
V = \begin{bmatrix} X_1 & X_2 & X_3 \\ 0 & X_2 & X_3 \\ 0 & 0 & X_3 \\ 0 & 0 & 0 \\
0 & 0 & 0 & X_3 \end{bmatrix}.
\]

These form Sierpinski gaskets.

We define a transformation \( \tilde{F} \) by
\[
V \tilde{F} U = \begin{bmatrix} 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ 0 & 0 & 0 & 1 \end{bmatrix},
\]
that is, \( V \tilde{F} U \) is the shift.

Then for a rectangular \( J \) which is a union of intervals cor-
responding to words such that its length of edges \( l_1, \ldots, l_d \)
satisfies \( l_1 \times l_2 \cdots \times l_d = 2^{-4d} \), we can show \( F^k(J) = L \).
From this fact, we can prove that the essential spectral ra-
dius of the Perron–Frobenius operator equals \( 2^{-d} \) and there
exists no eigenvalue except 1 in \( |z| > 2^{-d} \), thus the van der
Corput sequence generated by this transformation is of low
discrepancy. See for detail [2, 8].

References

[1] Yuko Ichikawa and Makoto Mor, “Discrepancy of
van der Corput sequences generated by piecewise lin-


Estimation of Beta-Value for pipelined beta encoders

Yeeai Chew and Yutaka Jitsumatsu

Department of Informatics, Kyushu University
744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan
Email: shu@me.inf.kyushu-u.ac.jp, jitumatu@me.inf.kyushu-u.ac.jp

Abstract—A β-converter is an analog-to-digital (A/D) encoder, that outputs truncated sequence of β expansion of an input value \( x \in [0, 1] \). β-converter has been proved to be robust to the fluctuation of the threshold value in quantizer. However, it remains an unsolved issue to give an accurate estimation of the β value in a pipeline β encoder. In this paper, we propose a new method estimating β by using β-map and the accuracy is also evaluated by numerical simulations.

1. Introduction

An A/D converter converts the continuous physical value to digital number and it is implemented in various of electronic equipment. Due to the large improvement in semiconductor microfabrication technology during recent years, A/D converters also tend to be more compact and having lower power consumption. Therefore circuit element’s value and threshold voltage play important roles and it is getting difficult to make sure of the conversion accuracy. For choosing an A/D conversion architecture, it is also important to concern about its electricity consumption, accuracy, conversion rate. Nowadays, i) Nyquist rate converter and ii) over sampled converter are the two main types of A/D converters which are commonly used.

Nyquist rate converter cuts off the signal whose frequency is over \( W \) by analog filter and samples the signal in the frequency over \( 2W \). After that, these sampled-value are converted into binary digits by the A/D converter. Among the Nyquist rate converter, the most popular one is natural weighted binary encoder also called PCM (Pulse Code Modulation) which gains the dyadic expansion of an input value \( x \in [0, 1] \). Although PCM is known to be easily calculated and achieves a precision of order \( O(2^{-N}) \) (\( N \) is bitrate), it makes error when the threshold voltage is fluctuated. On the other hand, \( \Sigma \Delta \) modulation is the typical over sampled converter. It has a self-correction property to the fluctuation of threshold by over-sampling in low quantization accuracy. \( \Sigma \Delta \) modulation is robust to the electric circuit elements and this property is the one that Nyquist rate converter does not own. Because of this robustness, we prefer \( \Sigma \Delta \) modulation to PCM while using imperfect quantizer, although \( \Sigma \Delta \) modulation owns slow conversion rate (achieves precision of order \( O(N^{-1}) \)).

β-converter is a new type of Nyquist rate converter proposed by Daubechies et al. in 2002 [1]. The most important fact about the β-encoder is that it is robust to the fluctuated quantizer while achieving a precision of order \( O(\beta^{-N}) \). β-encoder convert input-analog-signal \( x \in [0, 1] \) to digital bits by the expression

\[
x = \sum_{n=1}^{\infty} b_n \beta^{-n},
\]

where \( \beta \) is a real number satisfies the inequality \( 1 < \beta < 2 \) and \( b_n \in [0, 1] \). β-encoders have a look-up table (LUT) that memorizes the binary expansion of \( \beta^{-n} \) (\( n = 1, 2, \ldots, N \)). Such a LUT is used to convert the beta expansion coefficient \( \{ b_n \} \)s of \( x \) to binary expression of \( x \). β-encoder overcomes the disadvantage of both PCM and \( \Sigma \Delta \) modulation and it owns the potential to carry out A/D conversion in both high accuracy and speed.

When we mention about an β-encoder, there are two types of them. One is cyclic type which uses only one β-encoder to provide output bits and the other is pipeline β-encoder using plurality of β-encoder to provide outputs from each encoder. In order to apply faster analog digital conversion, it is important to construct pipeline β-encoder. As we can observed from expression (1), we need to know the exact β-value to restore the input value \( x \). We propose a new method for estimating β-value in pipeline β-encoder which uses β-map. Meanwhile, we also show the result of our numerical experiment estimating the β-value using the proposed method.

2. β-converter

2.1. cyclic-β-encoder

A β-encoder is composed of a β-times(\( 1 < \beta < 2 \)) amplifier and a quantizer with threshold value \( v \). β-encoder is known to be much more robust to the fluctuation of circuit elements than PCM. Moreover, β-encoder also could convert in higher rate than \( \Sigma \Delta \) converter, which means it can convert in both high rate and accuracy. Figure 2 shows a block diagram of β-encoder and when \( \beta = 2 \) it reduces to the PCM.

In a β-encoder, an input value \( x \in [0, 1] \) can be expanded into \( x = (\beta - 1) \sum_{n=1}^{\infty} b_n \beta^{-n}, b_n \in [0, 1] \). The expansion coefficients \( \{ b_n \} \) can be obtained as follows: we define a β-expansion map as

\[
C_\beta(x) = \begin{cases} \beta x, & x < v/\beta, \\ \beta x + 1 - \beta, & x \geq v/\beta. \end{cases}
\]

- 522 -
where \( \nu \in (\beta - 1, 1) \) is the threshold of the quantizer. Define a quantizer as
\[
Q_\nu(x) = \begin{cases} 
0, & x < \nu, \\
1, & x \geq \nu.
\end{cases}
\] (3)

Then \( b_n \) is given by the following expressions:
\[
\begin{align*}
\{ & u_1 = \beta x, b_1 = Q_\nu(u_1), \quad n = 0 \\
& u_{n+1} = \beta(u_n - b_n(\beta - 1)), b_{n+1} = Q_\nu(u_{n+1}), \quad n > 0 \}
\end{align*}
\] (4)

When we use a quantizer with error \( \epsilon \) in the threshold value \( \nu \), we can still carry out A/D conversion in a precision of order \( O(\beta^{-N}) \) if the threshold value remains in \((\beta - 1, 1)\), which means \( (\nu \pm \epsilon) \in (\beta - 1, 1) \). This characteristic is the one that PCM does not own, and so that \( \beta \)-encoder is said to be robust to the fluctuated quantizer.

2.2. \( \beta \)-estimation using Daubechies et al.’s method

In this section, a \( \beta \)-estimation method proposed by Daubechies et al. [2] is explained. It remains an important issue to know the exact value of \( \beta \) in order to conduct high accuracy A/D conversion using \( \beta \)-encoder. However, we cannot make exact \( \beta \)-converter because of the fluctuation of the circuit elements. Therefore it remains an important issue to estimate \( \beta \)-value in high accuracy after choosing a \( \beta \) times amplifier with errors (Although it seems there are two \( \beta \)-value remain estimating in Fig. 1, the \( \beta \)-value in two amplifier gain same value while using MDAC circuit [5]. So only one \( \beta \)-value remain estimating).

In terms of Daubechies et al.’s \( \beta \) estimating method, two input values \( x \in (0, 1) \) and \( 1 - x \in (0, 1) \) are used to produce \( L \) bits of \( \beta \)-outputs \( \{b_i\}_{i=1}^L \) and \( \{c_i\}_{i=1}^L \). Defining \( \gamma = \beta^{-1}, C = 1 + \nu + \epsilon, k_0 = \log(\frac{1+\epsilon}{1-\epsilon})/\log \gamma \), \( C' = \max\left(2C, 2C/(k_0y^{k_0-1})\right) \), the error of threshold value \( \nu \), then quantization error will satisfy the inequality
\[
0 \leq x - (1/\gamma - 1) \sum_{i=1}^L b_i \gamma^i \leq C' \gamma^L.
\] (5)

where \( \gamma = \beta^{-1} \). Meanwhile the quantization error of input value \( 1 - x \) also satisfies
\[
0 \leq 1 - x - (1/\gamma - 1) \sum_{i=1}^L c_i \gamma^i \leq C' \gamma^L.
\] (6)

From the Eqs. (5) and (6), we have
\[
0 \leq F_L(\gamma) \leq G_L(\gamma),
\] (7)
where \( F_L(\gamma) = 1 - (\frac{1}{\gamma} - 1) \sum_{i=1}^L (b_i + c_i)\gamma^i \) and \( G_L(\gamma) = 2C' \gamma^L \).

It can be easily verified that \( F_L(\gamma) \) is a monotone decreasing function of \( \gamma \) and that \( G_L(\gamma) \) is a monotone increasing function of \( \gamma \). According to these characteristics of \( F_L(\gamma) \) and \( G_L(\gamma) \), we can draw a graph in Fig. 2. From the inequality \( 0 \leq F_L(\gamma) \leq G_L(\gamma) \), we know that the true \( \gamma \) is limited in the region \( L \) illustrated in Fig. 2. Therefore when \( L \) is large enough, we could figure out the estimation value of \( \gamma \) expressed in \( \hat{\gamma} \) by finding the solution satisfies inequality(7) with Newton’s method or bisection method. After that, according to expression \( \gamma = \beta^{-1} \), estimated \( \beta \) value \( \hat{\beta} \) is calculated. In previous Oda’s research [3], they have also discovered a method finding the solution of inequality(7) by gradually enlarging \( L \). This method has a benefit that total number of calculation is decreased.\( \hat{\beta} \)'s are memorized in a LUT. Let the range of \( \beta \) be \( [\beta_{\min}, \beta_{\max}] \), which are divided into subintervals with equal width \( \Delta \beta \). The LUT memorizes \( \hat{\beta} \)'s for all candidates \( \hat{\beta} = \beta_{\min} + j \Delta \beta \), \( j = 0, 1, 2, ... \).

2.3. Pipeline \( \beta \)-encoder

Cyclic model and pipeline model are the two main models of the circuit structure when using Nyquist rate converter. If we assume getting \( L \) bits of output bits, the former one uses only one quantizer for \( L \) times, however, the latter one uses \( L \) pieces of quantizer to get the output bits. Consequently, the pipeline-model’s circuit area is \( L \) times larger than the cyclic-model’s but it could convert the signal in \( L \) times higher rate than the cyclic-model. We show the pipeline \( \beta \)-encoder in Fig. 3. For a pipeline \( \beta \)-encoder, let \( \beta_i \) be the amplification factor of \( i \)-th \( \beta \)-encoder (see Fig. 3). Then, we have
\[
x_i = \beta_{i-1}x_{i-1} - (\beta_i - 1)b_i \quad (i = 1, 2, 3, ..., L)
\] (8)

By using \( \beta \)-expansion in each stage, we can also obtain the following expression:
\[
x_L = \beta_{L-1}x_{L-1} - \cdots - (\beta_1 - 1)b_1 - (\beta_2 - 1)b_2
\]
From this expression, we could get the input value $x_0$ as the following expression since $x_L$ cannot be known.

$$x_0 = \sum_{i=1}^{L} \left( \beta_i - 1 \right) b_i \prod_{j=i+1}^{L} b_j.$$  (10)

Therefore, in order to reconstruct the input-value $x_0$ precisely from $b_i$, it is necessary to know the exact value of $\beta_i$.

3. $\beta$-estimation for pipeline-$\beta$ encoder

In section 2.2, we discussed about $\beta$-estimation method in cyclic $\beta$-encoder. However, in the case of pipeline $\beta$-encoder, we need to estimate plural $\beta$-value. While using Daubechies’s method estimating $\beta$-value, we can only achieve one inequality meaning impossible to estimate plural $\beta$. So that we need to find out a new method for estimating the $\beta$-value in pipeline $\beta$-encoder.

In this section, we explain our proposed method for estimating $\beta$-value for pipeline-$\beta$ encoder. In the following arguments, we suppose that:

- $\beta_1, \beta_2, ..., \beta_L$ are unknown (we could only know the range $[\beta_{\text{min}}, \beta_{\text{max}}]$)
- The dispersion of $\beta$-value tends to be larger in the later stage, however in this paper we supposed the dispersion does not differ in each stage.
- We can obtain output bits $b_1, ..., b_L$
- Input value $x_0$ cannot be made accurately and we also cannot get $x_1, ..., x_{L-1}, x_L$ directly

We update the $\beta$-value from the previously estimated value to a new one by the steps bellow:

1. We get as much input value $x_0$ as possible (suppose $K$ samples of $x_0$). $x_0^{(k)}$ means the $k$th input value for $k = 1, 2, ..., K$.

2. Let $b_1^{(k)} ... b_L^{(k)}$ represent the output bits for the initial value $x_0^{(k)}$ from $\beta$-encoder. The value $\beta_1, ..., \beta_L$ will be the target $\beta$ to estimate. We give up estimating the later $(L - L')\beta_i$, and considering their value as $\hat{\beta} = (\beta_{\text{min}} + \beta_{\text{max}})/2$. (In a practical implementation, one can use the nominal value of beta as $\beta$. However, in this simulation, we assume that the nominal value is given by $\hat{\beta} = (\beta_{\text{min}} + \beta_{\text{max}})/2$)

3. We estimate the $\beta$ in the order $\beta_{L'}, \beta_{L'-1}, ..., \beta_1$ (from the later one to the former one). The estimation procedure follows the steps below:

4-1 At this stage, we calculate pairs of reconstructed values ($\hat{x}_{i-1}^{(k)}, \hat{x}_i^{(k)}$) using output bits $b_1^{(k)} ... b_L^{(k)}$ according to the expressions:

$$\hat{x}_{i-1}^{(k)} = \sum_{n=i}^{L} (\hat{\beta}_n - 1) b_n^{(k)} \prod_{m=i+1}^{n} \beta_m^{-1}.$$  (11)

$$\hat{x}_i^{(k)} = \sum_{n=i+1}^{L} (\hat{\beta}_n - 1) b_n^{(k)} \prod_{m=i+1}^{n} \beta_m^{-1}.$$  (12)

Note that $\hat{\beta}_n$, the estimated value of $\beta_n$, have been already calculated at this moment for $n = i+1, i+2, ..., L$. Although in Eq. (11) it appears $\hat{\beta}_n$, which remain unconfirmed, we suppose its value as $\hat{\beta}$. Our aim is to estimate this $\hat{\beta}_i$.
of the right branch as \((\hat{x}_{i}^{(k_1)}, \hat{x}_{i}^{(k_1)})\). Then, we suppose that,  
\[
\theta = \frac{x_{i-1}^{(k_0)} + x_{i-1}^{(k_1)}}{2} \quad (14)
\]
\[
\hat{\beta}_i = \frac{\frac{x_{i}^{(k_0)}}{\theta} + 1 - \beta^{(k_1)}}{2} \quad (15)
\]

4.3 We update the estimated \(\hat{\beta}_i\) value according to Eqs. (14) and (15). Then, update \(\hat{x}_{i}^{(k_1)}\) again according to Eq.(10), after that repeat step (4-2). We quit updating \(\hat{\beta}_i\) after repeating a few times \(J\). Finally, let \(i \rightarrow i - 1\) and go back to 4-1 until we finish estimating all the \(\beta_i\).

4. Numerical Results

The accuracy of \(\beta\)-value is evaluated by computer simulation. In the simulation, we repeat the evaluation for 4 times while the number of input value \(x_0\) varies from \(10^5\), \(10^6\), \(10^7\) and the number of \(\beta\)-encoder are \(L' = 15\), \(L = 30\). We suppose the true \(\beta\) values are randomly selected in the range \(\beta \in [1.69, 1.71]\) according to the uniform distribution and start simulation from the initial value \(\hat{\beta} = 1.7\). The performances are evaluated by MSE (Mean Squared Error).

The simulation results show that the former \(\beta\)'s MSE is getting smaller than the later one. This is because the accuracy is gradually becoming accurate from the later \(\beta\) to the former \(\beta\). And comparing Fig. 5 and Fig. 6, the MSE is smaller as the input number \(K\) increase. From Fig. 7, we also know that when the repetition frequency increase, MSE does not decrease.

5. Conclusion

We proposed a new method estimating \(\beta\)-value in pipeline \(\beta\) encoder and conducted experiment through computer program. In our experiment, the MSE decrease comparing to the original error which implies the availability of our proposed method. We would like to implement our method in a electric circuit and verify its performance in the future work.
A Switched-Current Golden Ratio Encoder Circuit

Takayoshi Fujino† and Yoshihiko Horio††

†Graduate School of Engineering, Tokyo Denki University
5 Senju Asahi-cho, Adachi-ku, Tokyo 120-8551, Japan
Email: 15kme28@ms.dendai.ac.jp
††Research Institute of Electrical Communication, Tohoku University
2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

Abstract—A β-encoder is a novel analog-to-digital converter with a real number radix of β, which is robust against variations in circuit and environmental parameters. A golden ratio encoder (GRE) is a special case of the β-encoder where $\beta = \phi$ (golden ratio). This paper presents a GRE circuit based on the switched-current circuit technique. The proper operation of the proposed circuit is confirmed through numerical and SPICE circuit simulations.

1. Introduction

Digital circuits benefit from high density, high speed, and low power consumption and the CMOS process allows such circuits to be shrunk to a nanometer scale. On the other hand, as an interface between the analog and digital domains analog-to-digital converters (ADCs) are difficult to realize with high accuracy due to low-supply-voltage, low intrinsic gain of the advanced microfabrication transistors, and variations in component characteristics. Therefore, for example, an ADC architecture using a comparator without an operational amplifier has been proposed. However, in general, for high-performance ADC implementation, it is necessary to secure the accuracy of elements, and to introduce error correction techniques such as a digital compensation.

To alleviate these problems, a β-encoder has been proposed, which converts a real number to a digital bit sequence by using a real number radix based on the β-expansion [1, 2]. One of the advantages of the β-encoder is robustness against variations in component characteristics, temperature, and noise [3, 4]. Therefore, even with circuit elements with poor characteristics in an advanced nanometer semiconductor fabrication process, it is possible to realize an accurate ADC by exploiting this robustness.

The golden ratio encoder (GRE) is a special case of the β-encoder in which the radix value is the golden ratio. It is necessary to estimate the effective radix from the output bit sequence for the β-encoder. In contrast, we can always use the fixed radix $\phi = (1 + \sqrt{5})/2$ for the GRE. In addition, the GRE inherits all the advantages of the β-encoder.

In this paper, we propose a circuit implementation technique for the GRE. Because we aim at an integration of the GRE circuit with a low-voltage advanced microfabrication process, we employ the current-mode switched-current circuit technique. In addition, we confirm the functionality of the proposed circuit with SPICE circuit simulations and numerical simulations.

2. The Golden Ratio Encoder

The conversion algorithm of the GRE is given in the following:

$$\begin{align*}
\hat{u}_n &= \lambda_1 u_{n-1} + \lambda_2 p_{n-2} - \beta_0, \\
\beta_n &= Q_v(\hat{u}_{n-2}, \hat{u}_{n-1}).
\end{align*}$$

$$Q_v(\hat{u}_{n-2}, \hat{u}_{n-1}) = \begin{cases} 
1, & \hat{u}_{n-2} + \alpha \hat{u}_{n-1} < v, \\
1, & \hat{u}_{n-2} + \alpha \hat{u}_{n-1} \geq v.
\end{cases}$$

where $x_{\text{input}} \in [-1, 1]$ is the input value, $u_{-1} = x_{\text{input}}$ and $u_0 = 0$ are the initial values, $\alpha$ is the coefficient, $n$ is an integer time-index, $v$ is the threshold value of the quantizer, $\beta_n \in [0, 1]$ is the output bit on time $n$, and $u_n$ is the internal state on time $n$. As shown in Fig. 1, $\lambda_1$ and $\lambda_2$ are the non-unity coefficients considering the non-ideal transfer characteristics of the delay circuits. On the other hand, the decoded analog value $\hat{x}_{\text{input}}$ from the output bit sequence can be expressed as

$$\hat{x}_{\text{input}} = \sum_{n=0}^{L} b_n \phi^{-n},$$

where $L$ is the conversion bit length.

As noted above, GRE is based on the β-encoder, and therefore, can be derived from the equation of the β-encoder with $\beta = \phi$ [3]. Taking $\phi^2 = \phi + 1$, it is possible to perform the real number expansion of the radix $\phi$ without using any multiplier unit to generate $u_n$. In addition, calculation of the internal state value can be constructed only with unit delay elements.
2.1. Allowable range of threshold value $\nu$ and coefficient $\alpha$

Because GRE is a conversion method based on the $\beta$-encoder, variation is allowed if the threshold value $\nu$ and the coefficient $\alpha$ are within a certain range. If the allowable threshold values is $\delta$, i.e., $|\nu| \leq \delta$, and the coefficients $\lambda_1$ and $\lambda_2$ are $\lambda \in [0.9, 1]$, then the minimum and maximum allowable values of $\alpha$ are given by

$$\alpha_{\text{min}} = \frac{0.854(1 + \delta)}{0.618\lambda^2 + 0.236\lambda},$$

$$\alpha_{\text{max}} = 9.47 - \frac{6.472}{\lambda}(24.8\lambda^2 - 34.27\lambda + 11.71).$$

3. Switched-Current Golden Ratio Encoder Circuit

A block diagram of the proposed switched-current GRE circuit is shown in Fig. 2. As shown in the figure, each conversion stage takes the internal state values of the previous one and two times from the other conversion stage. Then, it outputs a single digital bit, and the internal state values of the previous one and current times to the other stage as a loop. The output bit sequence $b_1^1, b_2^1, \ldots, b_L^1$, and $b_L^2$ is obtained by repeating this loop, that is, [Stage 1 $\square$ Stage 2 $\square$ Stage 1 $\square$ $\cdots$] as many times as the bit length $L$. Figure 3 shows the conversion stage in Fig. 2, where $k = 1$ and 2. As shown in the figure, each stage is composed of a weighted 2-input adder, a quantizer, and a 3-input adder.

![Figure 2: Block diagram of the switched-current GRE circuit.](image)

![Figure 3: Block diagram of Stages 1 and 2 in Fig. 2.](image)

3.1. Weighted adder circuit

The weighted adder, which is surrounded by the solid line in Fig. 3, is realized by the circuit shown in Fig. 4. In addition, Fig. 5 defines the circuit symbol of this circuit. In the circuit of Fig. 4, the relationship between input voltages $V_{in1}$, $V_{in2}$, and $V_{in3}$, and output voltages $V_{OP}$ and $V_{OM}$ are given by Eq. (7) [7].

$$V_{OP} - V_{OM} = \frac{V_{AI}}{V_{AO}} \left( \frac{W_1}{L_1} (V_{in1} - V_{in}) + \frac{W_2}{L_2} (V_{in2} - V_{in}) \right),$$

where $K_p = \mu_pC_{vv}$ is the transconductance parameter of a $p$-type MOSFET. In addition, $V_{AI}$ and $V_{AO}$ are given by

$$V_{AI} = V_{in-cm} - V_{dm},$$

$$V_{AO} = V_{DD} - V_{out-cm} - |V_{hp}|,$$

where $V_{in-cm}$ is the input common mode voltage, $V_{out-cm}$ is the output common mode voltage, and $V_{dm}$ and $V_{hp}$ are the threshold voltages of $n$-type and $p$-type MOSFETs, respectively.

In the circuit shown in Fig. 4, when $L_1 = L_2 \equiv L$, $V_{in1} = V_{in2} \equiv V_{in}$, and the threshold voltage $\nu$ of GRE is $\nu = V_{in}$, the coefficient $\alpha$ in Eq. (3) can be determined by $W_1$ and $W_2$, namely, $\alpha = W_1/W_2$. Furthermore, the coefficient $\alpha$ can be close to the golden ratio by taking two subsequent numbers from the Fibonacci sequence, that is, 1, 1, 2, 3, 5, 8, $\ldots$, so that the aspect ratios ($W/L$) of MOSFETs M1 and M2 can be integers to approximate the value of $\phi$.

![Figure 4: The weighted adder circuit in Fig. 3.](image)

![Figure 5: The circuit symbol for the weighted adder circuit of Fig. 4.](image)
3.2. Golden ratio encoder circuit implementation

We propose a GRE circuit in Fig. 6 with the switched-current technique based on the CMOS analog inverter circuit. WA1 and WA2 in Fig. 6 correspond to the weighted adder circuits of Fig. 4. In addition, CP1 and CP2 are the latched comparators to implement the quantizer $Q(x, y)$ in Fig. 3.

In the switched-current circuits, capacitors for holding the sample values can be realized by the parasitic capacitances of the inverters. Therefore, the proposed GRE circuit makes it possible to miniaturize the circuit because it is mostly composed of MOSFETs without additional capacitors. Moreover, a switched-current circuit is suitable for low-power-supply voltage operation because of its small voltage swings at low-impedance nodes.

Figure 7 shows clock waveforms for driving the circuit in Fig. 6. We explain the circuit operation of Fig. 6 on the clock waveforms of Fig. 7. First, at time $t_0$, an input current signal is converted into voltage by inverter A1, and the current from $V_{in-cm}$ is by A2. These voltage values are held on the parasitic capacitances $C_{gs1}$, $C_{gs2}$, and $C_{gs3}$ of inverters A3, A4, and A5, respectively. At the same time, the converted voltage signals are applied to the weighted adder WA1. At time $t_1$, $b_1^1$ is outputted by the comparator CP1, and the outputs of Stage 1 are inputted to Stage 2. The operation of Stage 2 is similar to that of Stage 1, where the input current signals are converted into voltages by inverters A6 and A7, and they are held on parasitic capacitances $C_{gs4}$, $C_{gs5}$, and $C_{gs6}$ of inverters A8, A9, and A10, respectively. At the same time, these voltage signals are applied to the weighted adder WA2. At time $t_2$, $b_2^2$ is outputted by the comparator CP2, and the outputs of Stage 2 are inputted to Stage 1 recurrently. This recurrent process is repeated $L$-times, where $L$ is the bit length of the ADC.

4. Simulation Results

The operation of the proposed circuit was confirmed by numerical simulations and SPICE circuit simulations. In the simulations, we evaluate the return maps of the internal state, and the encode–decode characteristics.

4.1. Return maps

Figure 8 shows the return map of $u_{n-1} - u_{n-2}$ obtained by numerical simulation of Eqs. (1)–(3) with $\alpha = 1.6$ and $\nu = 0$. Figure 9 shows the corresponding SPICE simulation results of the proposed circuit using ideal circuit elements. From the results of Fig. 8 and Fig. 9, it can be confirmed that the overall behavior of the internal state obtained from the SPICE simulation agrees with that from the numerical simulation.

4.2. A/D–D/A conversion characteristic

Figure 10 shows the A/D–D/A conversion (encode and decode) characteristic by the numerical simulation of
Figures 8 and 9 show the return maps of $u_{n-1}$–$u_{n-2}$ (Numerical simulation) and (SPICE simulation) respectively. Figure 10 and Figure 11 show the A/D–D/A conversion characteristics (Numerical simulation) and (SPICE simulation) respectively.

Eqs. (1)–(4) with $x = 1.6, v = 0$, and $L = 12$. Figure 11 shows the corresponding SPICE simulation results of the proposed GRE circuit. From these figures, it can be confirmed that the encode–decode characteristic of the SPICE simulation agrees with that of the numerical simulation. Hence, it is confirmed that the proposed switched-current GRE circuit faithfully implements the GRE equations.

5. Conclusion

We have proposed the GRE circuit based on the switched-current circuit technique with the analog CMOS inverter circuit. The operation of the proposed circuit was confirmed by numerical and circuit simulations.

Acknowledgment

This work is supported by VLSI Design and Education Center (VDEC), the University of Tokyo in collaboration with Cadence Design Systems, Inc.

References

Parameter Optimization for Power Line Communications Considering Operational Status of Electrical Appliances

Takeshi Kamio†, Ena Kono†, and Hisato Fujisaka†

†Faculty of Information Sciences, Department of Systems Engineering, Hiroshima City University
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima, 731-3194, Japan
Email: kamio@hiroshima-cu.ac.jp

Abstract— Power line communication (PLC) is considered as one of communication systems to support a smart grid. Especially, PLC has the advantage that the construction cost is very low, since it can use existing power lines. G3-PLC is an international standard for the low-frequency narrow-band OFDM PLC and it is adopted in Japan. OFDM (orthogonal frequency division multiplexing) is a multi-carrier modulation scheme. Since the transmission-line impedance and noise vary with the operational status of electrical appliances, both the primary modulation and the allocated power for each carrier should be optimized to maximize the transmission capacity.

In this paper, we assume that the transmission-line impedance and noise are obtained by the electrical appliance monitoring system which is incorporated in the smart meter. Under this condition, we formulate the transmission capacity, whose variables are the primary modulation and the allocated power for each carrier. Also, we attempt to optimize these variables to maximize the transmission capacity by PSO (particle swarm optimizer). Finally, the usefulness of our proposed method has been confirmed by numerical experiments.

1. Introduction

A smart grid is a next-generation electrical power grid which can optimize the power balance between supply and demand by using the information communication technology (ICT). Also, since it helps to efficiently use the electric energy and introduce the renewable energy such as solar power, it is regarded as one of measures against global warming.

A smart meter is an electric power meter equipped with communication capabilities. Therefore, it is an important device which combines the consumer with the smart grid. There are two communication routes, called routes A and B, whereby data is directly obtained through the smart meter. Route A links smart meters with electric power companies and route B links smart meters with HEMSs (home energy management systems) [1]. There are plural candidates for communication systems applied to routes A and B. From the viewpoint of the construction cost, PLC (power line communication) [2] has attracted a lot of attention.

In the case of Japan, the communication system for route A is selected from wireless multi-hop system, one-to-many (1:N) wireless system, and PLC system. On the other hand, the communication system for route B is selected from 920MHz wireless system and PLC system. Especially, when PLC system is used for route B, the PLC must keep G3-PLC which is an international standard specification for the low-frequency narrow-band OFDM PLC. OFDM is a multi-carrier modulation scheme.

From these backgrounds, it is clear that PLC is a key technology supporting the smart grid. However, since the transmission-line is the power line, the transmission-line impedance and noise (i.e., characteristics of transmission-line) vary with the operational status of electrical appliances. Therefore, to maximize the transmission capacity of OFDM PLC (e.g., G3-PLC), both the primary modulation and the allocated power for each carrier must be optimized according to the characteristics of transmission-line.

In this paper, to satisfy the above request, we propose the following method. It is assumed that the information about both the transmission-line impedance and noise are obtained by the electrical appliance monitoring system [3] which is incorporated in the smart meter. Under this condition, we formulate the transmission capacity, whose variables are the primary modulation and the allocated power for each carrier, considering the transmission-line impedance and noise. Also, we attempt to optimize these variables to maximize the transmission capacity by PSO (particle swarm optimizer) [4]. Finally, the usefulness of our proposed method has been confirmed by numerical experiments.

2. Model of Power Line Communications

Fig.1 shows a circuit of typical PLC system. For example, if a HEMS (Home Energy Management System)
requests data on the electric power consumption to a smart meter, the smart meter is a transmitter and the HEMS is a receiver in route B. In this section, we consider the influence of the transmission-line impedance and noise on PLC.

As shown in Fig.1, when the power supply voltage (e.g., 100VAC, 50/60Hz in Japan) is given by \( V_S \), the transmitting signal voltage (i.e., data) of a carrier, which is restricted within the PLC band (e.g., 10k~450kHz in Japan), is given by \( V_r \), the composite voltage \( V_c = V_S + V_r \) is applied to the power line. Although only the signal corresponding to PLC band can pass through the filter, the received signal voltage \( V_R \) is not equal to \( V_r \). Because \( V_r \) is changed to \( V_R \) by the transmission-line impedance and noise which vary with the operational status of electrical appliances. Considering these facts, we derive the relationship between the output voltage \( V_o \) of the receiver and \( V_R \) from the equivalent circuit shown in Fig.2. In this figure, \( Z_T \) and \( Z_A \) are impedances of the transmitter and receiver respectively. \( Z_d \) is the combined impedance of electrical appliances in operation. \( V_{RN} \) and \( V_{AN} \) are the noise generated by the receiver and electrical appliances in operation respectively. Moreover, \( A \) is the voltage gain of receiver. Therefore, \( V_o \) is given as follows:

\[
V_o = \frac{A Z_a Z_T / Z_s V_r}{Z_t + Z_s / Z_R} + \frac{A Z_T / Z_s V_{RN} + V_{RN}}{Z_t + Z_s / Z_R}. \tag{1}
\]

This equation shows that the carrier to noise ratio (C/N) is given by \( C/N = \alpha^2 / \beta \), if \( V_r \) is given by a sinusoidal voltage. Therefore, the BER (bit error rate) of each carrier in the PLC band can be calculated by substituting \( C/N \) for the theoretical formula [5].

In this research, we assume that the electrical appliance monitoring system [3] can recall \( Z_d \) and \( V_{AN} \) for the known operational status of electrical appliances and can evaluate them for the unknown one. Therefore, the larger the operating time of the monitoring system becomes, the more easily \( \alpha \) and \( \beta \) in Eq.(1) can be obtained.

3. Parameter Optimization for OFDM PLC

3.1. Cost Function

To maximize the transmission capacity of PLC using OFDM as the secondary modulation (i.e., OFDM PLC), both the primary modulation and the allocated power for each carrier must be optimized according to the characteristics of transmission-line. To satisfy this request, we define the cost function which consists of one objective and two constraint terms.

The objective term \( f_1 \) is derived from the transmission capacity \( R \). Here, we assume that the \((m, n)\)-Hamming code (HC\((m, n)) \) is used as the error correcting code (ECC) and the primary modulation for each carrier is selected from M-QAM (\( M \in \{4, 16, 64, 256, 1024\} \)). If the transmission capacity \( R \) is defined as the number of information data bits which are conveyed per unit of time and also demodulated without an error, then \( R \) is given by

\[
R = \sum_{i=1}^{N_C} \frac{B_i}{T_i} \frac{n}{m} \frac{p_{\text{ref}}}{\gamma}, \tag{2}
\]

where, \( k \) is the carrier index, \( N_C \) is the number of carriers, \( B_i \) is the number of bits per symbol decided by the primary modulation, \( T_i \) is the symbol length, \( n/m \) is the coding rate of HC\((m, n)) \), and \( p_{\text{ref}} \) is the probability that HC\((m, n)) \) correctly extracts \( n \) bits information data from \( m \) bits received data, in other words, the probability that \( m \) bits received data is demodulated without an error. If there is only one bit error in \( m \) bits received data, HC\((m, n)) \) can correct the error. Therefore, \( p_{\text{ref}} \) is given as follows:

\[
p_{\text{ref}} = C_i \left(1 - \left(1 - p_{\text{ref}}\right)^m\right)^i + \left(1 - \left(1 - p_{\text{ref}}\right)^m\right)^i \left(1 - (m-1)p_{\text{ref}}\right), \tag{3}
\]

where, \( p_{\text{ref}} \) is the BER (bit error rate). In this research, we assume that PLC is under the additive white Gaussian noise (AWGN) channel and Gray code bit mapping is employed. Therefore, the BER of \( k \)-th carrier using M-QAM is calculated by substituting \( C/N_k \) for the following equations [5]:

\[
p_i (a) = \frac{1}{\sqrt{M}} \sum_{\omega = 0}^{M-1} e^{2\pi i a \omega / M} \left[ \frac{1}{\sqrt{M}} \left[ 2^{\omega - i \frac{1}{M} - i \frac{1}{2}} \right] \right], \tag{4}
\]

\[
p_i = \frac{1}{\log_2 \sqrt{M}} \sum_{i=1}^{M} p_i (a). \tag{5}
\]

The objective term \( f_1 \) is defined by the transmission capacity \( R \) and its reference value \( R_{\text{ref}} \) as follows:

\[
f_1 = \ln \left( \frac{R}{R_{\text{ref}}} \right). \tag{6}
\]

If \( R \) becomes smaller than \( R_{\text{ref}} \), \( f_1 \) decreases rapidly.

Next we describe two constraint terms (\( f_2 \) and \( f_3 \)). The first constraint term \( f_2 \) is necessary to guarantee that HC\((m, n)) \) correctly extracts \( n \) bits information data from \( m \) bits received data. As mentioned above, if there is only one bit error in \( m \) bits received data, HC\((m, n)) \) can correct the error. Therefore, if the reference BER \( p_{\text{ref}} \) is set to \( 1/m \), the BER of \( k \)-th carrier \( p_k \) must satisfy \( p_k \leq p_{\text{ref}} \). To satisfy this condition at all the carriers, the constraint term \( f_2 \) is defined as follows:

\[
f_2 = \sum_{i=1}^{N_C} U(p_k - p_{\text{ref}}) \ln \left( \frac{p_{\text{ref}}}{p_k} \right), \tag{7}
\]

where, \( U(x) \) is the unit step function. Also, when the \( k \)-th carrier is not used, \( p_k \) is set to zero. Therefore, even if only one carrier satisfies \( p_k / p_k < 1 \), \( f_2 \) decreases rapidly. On
the other hand, if all the carriers satisfy $p_{ct}/p_k \geq 1$, $f_2$ becomes zero.

The second constraint term $f_3$ is necessary to guarantee that the total power consumption $E_T$ of all the carriers is smaller than or equal to the reference value $E_{ct}$. To satisfy this condition ($E_T \leq E_{ct}$), the constraint term $f_3$ is defined as follows:

$$f_3 = U(E_T - E_{ct}) \ln \left( \frac{E_{ct}}{E_T} \right).$$  \hspace{1cm} (8)

Therefore, if the PLC system consumes the electric power more than $E_{ct}$, $f_3$ decreases rapidly. On the other hand, if $E_T \leq E_{ct}$, $f_3$ becomes zero.

Using the objective term ($f_1$) and two constraint terms ($f_2$ and $f_3$), the cost function $f$ is defined as follows:

$$f = y_1 + (1-\gamma) y_2 + (1-\gamma)(1-\delta) y_3,$$  \hspace{1cm} (9)

where, $\gamma (0 < \gamma < 1)$ is a weight coefficient which controls the relationship between the objective term and two constraint ones. Also, $\delta (0 < \delta < 1)$ is a weight coefficient which controls the relationship between two constraint terms.

### 3.2. Parameter Optimization by PSO

In this paper, we attempt to optimize both the primary modulation and the allocated power for each carrier by applying PSO to maximize the cost function in Eq.(9). As mentioned above, $B_k$ is the number of bits per symbol decided by the primary modulation applying to the k-th carrier. The primary modulation is selected from M-QAM ($M \in \{4, 16, 64, 256, 1024\}$). Moreover, we assume the probability that the k-th carrier is not used to convey the data. Therefore, $B_k$ corresponds to one of elements in the set $\{0, 2, 4, 6, 8, 10\}$. Considering these conditions, we define the position vector $x \in \mathbb{R}^{2N_e}$ of each particle as follows:

$$x = \left( \{b_1, E_1\}, \ldots, \{b_x, E_x\}, \ldots, \{b_N, E_N\} \right),$$  \hspace{1cm} (10)

where, $b_k$ satisfies $0 \leq b_k \leq 10$ and $B_k$ is an element in the set $\{0, 2, 4, 6, 8, 10\}$ which is the closest to $b_k$. If $B_1$ is zero, the k-th carrier is not used. On the other hand, if $B_1$ is not zero, the primary modulation is 2$^B_k$-QAM. Also, $E_k$ is the allocated power for k-th carrier and satisfies $0 \leq E_k \leq E_{ct}$. If $B_1$ is zero, $E_1$ is set to zero forcibly.

The PSO model used in this research is composed of the original PSO developed by J. Kennedy et al. [4] and the reset function. We call it O-PSO-R (O-PSO with reset). In the case of O-PSO-R, each particle keeps the personal best position $x_{TB}$ and its evaluation value $f_{TB}$, and the swarm keeps the global best position $x_{GB}$ and its evaluation value $f_{GB}$. However, the O-PSO-R eliminates the information about personal/global best at the time of the reset. To avoid the loss of the best solution during the searching process, O-PSO-R stores the best position and its evaluation value at $x_{TB}$ and $f_{TB}$ respectively. The subscript "TB" means the trial best. Although the trial best position $x_{TB}$ is not used in the update equations in the PSO algorithm, it is used to re-initialize the position of each particle in the reset process. However, O-PSO-R does not guarantee that $x_{TB}$ is a feasible solution. Therefore, if $x_{TB}$ is a feasible solution, $x_{TB}$ and $f_{TB}$ are copied at $x_{TB}$ and $f_{TB}$ as the best feasible solution. Also, O-PSO-R has several special processes in the initialization, the judgement of the reset, the re-initialization, and the end condition as follows.

At the start of the search by O-PSO-R, the position of each particle is initialized based on the feasible solution $x_{SM}$ obtained by the simple mapping method shown in Fig.3. Concretely, a particle is set at the position $x_{SM}$ and others are set at the positions given by adding the random numbers to $x_{SM}$. The random numbers added to $b_k$-components are given by $r_0 = [-0.5, 0.5]$. On the other hand, the random numbers added to $E_k$-components are given by $r_E = [-E_{ref}/40, E_{ref}/40]$. Moreover, the velocity of each particle is initialized by randomly choosing the numbers from $r_v$ and $r_E$.

The purpose of the reset is to improve TB(i.e., $x_{TB}$ and $f_{TB}$). Therefore, if the probability of the improvement of TB is low, the reset should be executed. O-PSO-R has the following four conditions to judge the execution of reset.

1. **After the initialization/re-initialization, it is assumed that** $f_{TB}$ **has not been improved. Also, if the improvement of** $f_{GB}$ **per iteration is less than 1.0e-6 for 100 successive iterations, the reset is executed.**

2. **After the initialization/re-initialization, it is assumed that** $f_{TB}$ **has been improved. Also, if the improvement of** $f_{GB}$ **per iteration is less than 1.0e-12 for 100 successive iterations, the reset is executed.**

3. **After the initialization/re-initialization, it is assumed that** $f_{TB}$ **has not been improved. If the number of iterations achieves 1000, the reset is executed.**

4. **After the initialization/re-initialization, it is assumed that** $f_{TB}$ **has been improved. After the improvement of** $f_{TB}$, **the number of iterations achieves 1000, the reset is executed.**

At the time of the reset, the position of each particle is re-initialized based on $x_{TB}$. All the particles are set at the positions given by adding the random numbers to $x_{TB}$ in the same way as the initialization at the start of the search. Also, the velocity of each particle is re-initialized by randomly choosing the numbers from $r_v$ and $r_E$.

There are two end conditions of a searching trial. The first condition is that the number of resets after the last improvement of $f_{TB}$ achieves 3. The second condition is that the total number of iterations achieves 20000. When one of them is satisfied, O-PSO-R stops searching.
4. Numerical Experiments

Numerical experiments have been carried out by using the measured data to evaluate our proposed method. As mentioned above, our method can optimize the primary modulation and the allocated power for each carrier to maximize the transmission capacity of OFDM PLC considering both transmission-line impedance and noise. Unfortunately, since we could not obtain the measured data of transmission-line impedance, the following experiments use only the measured data of noise.

The conditions of OFDM PLC are as follows. The band ranges from 11 to 40kHz and the carrier spacing $\Delta f$ is 1kHz. Therefore, the symbol length $T_S$ is given by $T_S=1/\Delta f =1$ms and the number of carriers $N_C$ is 30. Also, since HC(7, 4) is used, the coding rate $n/m$ is 4/7 and the reference BER ($p_{ref}$) is set to 1/7. Moreover, the reference value of power consumption $E_{ref}$ is 1.0e-5 and the value of transmission capacity $R_{ref}$ is set to $R_{SM}$ which is given by the simple mapping shown in Fig.3. The conditions of O-PSO-R are as follows. The weight coefficients of cost function is given by $\gamma=\delta=0.5$. The number of particles is 30. The inertia weight coefficient $w$ and acceleration coefficients $c_1, c_2$ included in the update equation of PSO are given by $w=0.729$, $c_1=c_2=1.49445$ respectively. The number of searching trials is 20. The other conditions are as mentioned in Sect. 3.

Table 1 shows the performance comparisons between the simple mapping method and our proposed method (i.e., O-PSO-R). From the viewpoint of the maximization of the transmission capacity, it has been confirmed that O-PSO-R is much superior to the simple mapping. Moreover, we show the solutions of the simple mapping and O-PSO-R with the best $R$ in Fig.4 and Fig.5 respectively. From these results, it has been found that our method can change the some primary modulations decided by the simple mapping for better ones by controlling the allocated power.

5. Conclusions

In this paper, we have proposed a PSO-based method to optimize the primary modulation and the allocated power for each carrier to maximize the transmission capacity of OFDM PLC. From numerical experiments, we have confirmed that our O-PSO-R can improve the feasible solution given by the simple mapping method.

In the future, we will have to conduct more experiments using measured data of not only the noise power but also the transmission-line.

References

Piecewise-linear particle swarm optimizer networks

Tomoyuki Sasaki†‡, Hidehiro Nakano†, Arata Miyauchi† and Akira Taguchi†

†Department of Computer Science, Tokyo City University, 1-28-1, Tamazutsumi, Setagaya-ku, Tokyo, 158-8557, Japan
‡Research Fellow of Japan Society for the Promotion of Science Email: sasaki@ic.cs.tcu.ac.jp

1. Introduction

Particle swarm optimizer (PSO) [1] which has been developed by Kennedy and Eberhart in 1995 is one of the population-based stochastic algorithms. PSO mimics social behaviors of creatures such as birds flocking and fish schooling. These creatures are represented by particles as solution candidates, which search a multi dimensional search space and find feasible solutions. PSO has following advantages of: (1) relatively few control parameters; (2) that it is easy to implement PSO to applications; and (3) quick convergence characteristics. Therefore, PSO is applied to various applications.

In order to find good feasible solutions for large scale optimization problems, it is required that many particles search a search space. However, calculation costs are increased in proportion to the number of particles. The calculation costs of particles can be distributed by using multiple PSO circuits in parallel, which calculate the behavior of each particle in the original PSO [2]; however, the PSO circuit has random number generators and multiple floating point multipliers, because the original PSO particle has the stochastic factors. As such, the circuit amount of the original PSO becomes large, and it is hard to implement large number of the circuits on hardware. Therefore, decreasing the circuit amount is required.

In our previous study, piecewise-linear particle swarm optimizer (PPSO) [3] which is one of the deterministic PSOs has been proposed. PPSO particle has two dynamics, which are the convergence and the divergence modes, and searches a search space by switching both dynamics. The solving performances of PSPO are substantially same as those of the original PSO. Furthermore, since PSPO does not have stochastic factors, it can be realized that the size of a PPSO circuit is smaller than that of the PSO circuit [2]. In order to solve large scale optimization problems, a large number of PPSO circuits is required. However, a parallel computing method of PPSO circuits has not been studied.

PSO Network (PSON) [4], which is one of the parallel computing methods for PSO, has been proposed. In PSON, a population is divided into multiple sub-PSOs, and each sub-PSO is connected to neighbor sub-PSOs via network structure. Each sub-PSO searches a search space independently, and communicates own best solution to the neighbor sub-PSOs. When each sub-PSO is assigned by single processor, evaluation costs of the population can be distributed. Furthermore, PSON has better solving performances than the original PSO.

In this paper, we propose a model of networking PPSO (PPSON) for parallel computing. In PPSON, the concept of PSON is applied to PPSO; a population of PPSO is divided into multiple sub-PPSOS, and each sub-PPSO searches a search space independently. Each sub-PPSO is connected to neighbor sub-PPSOS which are determined via network structure, and communicates own best information to the neighbor sub-PPSOS. The effectiveness of PPSON is investigated by numerical experiments compared with PSON and PPSO.

2. Piecewise-linear Particle Swarm Optimizer (PPSO) [3]

In this section, the basic idea of a piecewise-linear particle swarm optimizer (PPSO) is explained. The $i$th particle has velocity vector $v_i = (v_{i1}, v_{i2}, \ldots, v_{iD})$, position vector $x_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, and $pb_i = (pb_{i1}, pb_{i2}, \ldots, pb_{iD})$ which is the personal best solution, and shares $gb = (gb_1, gb_2, \ldots, gb_D)$ which is the global best solution in a swarm. $D$ denotes the number of design variables. Furthermore, PPSO particle has convergence and divergence modes.

The updating rules of the $j$th component of the $i$th particle in PPSO are described by

\[ q_{ij} = (1-r)pb_{ij} + rgb_j \]
\[ y_{ij} = x_{ij} - q_{ij} \]
\[ \begin{bmatrix} v_{ij}^{\text{new}} \\ y_{ij}^{\text{new}} \end{bmatrix} = \delta_{ij} \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix} \begin{bmatrix} v_{ij}^{\text{old}} \\ y_{ij}^{\text{old}} \end{bmatrix} \]

where $r$ denotes a constant parameter. $y_{ij}$ denotes a relative position from the equilibrium point $q_{ij}$ to the $i$th particle’s position $x_{ij}$. $\delta_{ij}$ denotes a damping factor, and $\theta$ denotes a
rotation angle. Each particle has two search modes, convergence mode and divergence mode, by switching the damping factor $\delta_{ij}$. The switching rule from the convergence mode to the divergence mode is given by the following.

$$\text{if } v_{ij}^{\text{old}} \cdot v_{ij}^{\text{new}} \leq 0 \text{ and } |y_{ij}^{\text{new}}| < T h_{i,j}$$

$$\begin{align*}
\delta_{ij} &= \delta_{d} \\
v_{ij}^{\text{new}} &= 0 \\
T h_{i,j}^{\text{new}} &= \alpha |y_{ij}^{\text{new}}|
\end{align*}$$

where $\delta_{d} > 1$ denotes a damping factor in the divergence mode. $T h_{c}$ denotes a switching threshold in the convergence mode, which is updated at the end of the convergence mode. $\alpha$ denotes a scaling parameter of $T h_{c}$.

On the other hand, the switching rule from the divergence mode to the convergence mode is given by the following.

$$\text{if } v_{ij}^{\text{old}} \cdot v_{ij}^{\text{new}} \leq 0 \text{ and } |y_{ij}^{\text{new}}| > T h_{d,j}$$

$$\begin{align*}
\delta_{ij} &= \delta_{c} \\
v_{ij}^{\text{new}} &= 0 \\
T h_{d,j}^{\text{new}} &= T h_{d,j} + \beta(T h_{c,j} - T h_{d,j})
\end{align*}$$

where $0 < \delta_{c} < 1$ denotes a damping factor in the convergence mode. $T h_{d}$ denotes a switching threshold in the divergence mode, which is updated at the end of the divergence mode. $\beta$ denotes a scaling parameter of $T h_{d}$.

The procedures of PPSO algorithm for minimizing objective function $f(x)$ are explained below.

**Step1: Initialization**

Set the maximum iterations $I_{\text{max}}$ and the total number of particles $N$. For all $i$, $x_{i}$ and $v_{i}$ are initialized at random. Let $t = 0$.

**Step2: Update the best solutions**

For all $i$, evaluate the fitness value of the $i$th particle and update $p_{b}^i$ by the following equation.

$$p_{b}^i = \begin{cases} x_{i} & \text{if } f(x_{i}) < f(p_{b}^i) \text{ or } t = 0 \\ p_{b}^{i-1} & \text{otherwise} \end{cases}$$

Then, update $g_{b}^i$ by the following equations.

$$k = \arg \min_i f(p_{b}^i)$$

$$g_{b}^i = p_{b}^k$$

**Step3: Update velocity and position vectors**

For all $i$, the velocity and position vectors of the $i$th particle are updated by Eqs. (3) ~ (7).

**Step4: Judgment of termination**

$t = t + 1$. Then, if $t \neq I_{\text{max}}$, go to Step 2.

The idea of PPSO is very simple and it is easy to implement PPSO on digital or analog circuit. As PPSO is implemented on an analog circuit, it can be realized by nonlinear resistor, capacitor, inductor and voltage sources.

3. Networks of PSOs and PPSOs

3.1. PSO Networks (PSON) [4]

PSON is one of the sub-swarm PSO methods. In PSON, a population is divided into multiple sub-PSOs. Each sub-PSO is connected to neighbor sub-PSOs which are determined by network structure. The $th$ sub-PSO searches a search space independently, and communicates the own local best solution $b_{g} = (b_{g,1}, b_{g,2}, \ldots, b_{g,N})$ among neighbor sub-PSOs. The $th$ sub-PSO has the group local best solution $g_{i} = (g_{i,1}, g_{i,2}, \ldots, g_{i,N})$ among the neighbor sub-PSOs. Particles update velocity and position by referring to $p_{b}$, $b_{g}$ and $g_{i}$. Updating rules of the $i$th component of the $j$th particle’s information in the $j$th sub-PSO are described by the following equations.

$$v_{g_{i,j}}^{t+1} = w v_{g_{i,j}}^{t} + c_{1} r_{1} (p_{b_{i,j}}^{t} - x_{g_{i,j}}^{t}) + c_{2} r_{2} (b_{g_{i,j}}^{t} - x_{g_{i,j}}^{t})$$

$$x_{g_{i,j}}^{t+1} = x_{g_{i,j}}^{t} + v_{g_{i,j}}^{t+1} + c_{3} r_{3} (g_{i,j}^{t} - x_{g_{i,j}}^{t})$$

The network topology of PSON is characterized by $DBG$, which denotes the degree of connection between sub-PSOs. Figure 1 shows examples of PSON. The number of sub-PSOs is 4 and the number of particles of each sub-PSO is 6 in these examples. In (a), each sub-PSO is connected to two neighbor sub-PSOs (Ring topology), while in (b), to all other sub-PSOs (Fully connected topology). In PSON, even if one sub-PSO converges to a local optimum solution, the sub-PSO can escape from the local optimum solution by referring to $g_{i}$. Furthermore, parallel computing can be realized by assigning single processor to each sub-PSO.

3.2. PPSO Network (PPSON)

In this section, we propose PPSO Network (PPSON). The networking method of PSON is applied to PPSO. A population of PPSO is divided into multiple sub-PPSOS which have the own best solution $b_{g}$ and group local best solution $g_{i}$, and each sub-PPSO is connected to neighbor sub-PPSOS. Each sub-PPSO searches a search space independently, and communicates own $b_{g}$ to the neighbor sub-PPSOS. In PPSON, there are cases where sub-PPSOS do not converge to $g_{i}$ if an equilibrium point $q$ is set by the gravity of $p_{b}$, $b_{g}$ and $g_{i}$. Because the information of $g_{i}$ involves the information of $b_{g}$, an equilibrium point of each sub-PPSO is defined as follows.

$$q_{ij} = (1 - r) b_{p_{ij}} + r g_{l_{j}}$$
Table 1: Simulation settings

<table>
<thead>
<tr>
<th></th>
<th>PPSON</th>
<th>PSON</th>
<th>PPSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of Groups (G)</td>
<td>10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>No. of Particles (N)</td>
<td>5</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>(DBG)</td>
<td>2.9</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Dimension (D)</td>
<td>50</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iterations (t_{\text{max}})</td>
<td>20000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trials</td>
<td>100</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where \(r\) denotes a constant parameter.

The procedures of PPSON algorithm for minimizing objective function \(f(x)\) are explained below.

**Step 1: Initialization**
Set the maximum iterations \(t_{\text{max}}\), the total number of sub-PPSOs \(G\), and the total number of particles \(N\) for each sub-PPSO. Let \(t = 0\).

For all \(i\), \(x_i^0\) and \(v_i^0\) are initialized at random, and \(p_{bf}^0\) is initialized by \(x_i^0\). For all \(g\), \(l_b^0\) and \(g_l^0\) are initialized by the best \(p_{bf}^i\) in PPSO, and \(g_l^0\) is initialized by \(l_b^0\).

**Step 2: Update velocity and position vectors**
For all \(i\), the velocity and position vectors of the \(i\)th particle are updated by Eqs. (3) ~ (7).

**Step 3: Update the best solutions**
For all \(g\), PPSO is evaluated. For all \(i\), the \(i\)th particle in PPSO is evaluated, and \(p_{bf}^g\) is updated. Then, \(l_b^g\) is updated by the best \(p_{bf}^g\).

**Step 4: Communication for each sub-PPSOs**
Perform synchronization of all sub-PPSOs at regular iterations which are decided by a constant parameter \(Period\), and each sub-PPSO sends the own \(l_b\) to neighbor sub-PPSOs. The updating rules of the \(g\)th sub-PPSO for \(g_l\) are described by the following equations.

\[
k = \arg \min_{j \in n(g)} f(l_{bf}^j)
\]

\[
g_l^g = l_b^k\]  

where \(n(g)\) denotes the \(g\)th sub-PPSO’s neighbor sub-PPSOs.

**Step 5: Judgment of termination**
\(t = t + 1\). Then, if \(t \neq t_{\text{max}}\), go to Step 2.

4. Numerical experiments

In order to confirm effectiveness of PPSON, PPSON is compared with PSON and PPSO. Table 1 and Table 2 show the simulation conditions and benchmark functions of which the optimal solution is 0, respectively. We adopt the combination of parameters which leads to the best result, shown in Table 3. In PPSON and PPSO, we set \(\delta_c = 0.65\), \(\delta_d = 1.75\), and \(\alpha = 1.0\). Initial \(T_{hd}\) is a maximum range of each benchmark function, because particles should search a search space globally in an initial search stage. These parameter conditions realize that each particle can converge to the equilibrium point.

Table 4 shows the simulation results. In Table 4, “Mean” denotes average fitness value, and “SD” denotes the standard deviation. As shown in Table 4, the solving performances of PPSON are better than those of the others for multimodal functions. While, in non-separable unimodal functions, PSON has better solving performances than PPSON, because PSON has better local search ability than PPSON. Since each sub-PPSO can search a better solution by communicating to neighbor sub-PPSOs, the solving performances of PPSON are improved more than those of PPSO. In addition, when each sub-PPSO is assigned by single processor, high parallelism can be realized. As such, the networking method for PPSON can realize that the solving performances are improved and the calculation costs are distributed. The solving performances of PPSON for fully connected topology \((DBG = 9)\) are better than for ring type topology \((DBG = 2)\). Therefore, each sub-PPSO should refer to the best information in all sub-PPSOs. In addition, \(r\) should be set suitable value for problems.

5. Conclusion

In this paper, we proposed the model of networking PPSO for parallel computing. PPSO was applied to the concept of PSON, and the solving performances of PPSON were compared with PSON and PPSO. The simulation results showed that PPSON has better solving performances than PSON and PPSO for multimodal problems. In addition, when each sub-PPSO is assigned by single processor, high parallelism can be realized. Therefore, it was clear that PPSON is effective for parallel computing.

In our future works, we would like to implement PPSON algorithm on digital or analog circuit to solve engineering optimization problems.
Table 3: Parameter settings for each method

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter Settings</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>Period = 10</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.7</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Rosenbrock</td>
<td>Period = 20</td>
<td>0.85</td>
<td>0.65</td>
<td>0.7</td>
<td>1.0</td>
<td>0.9</td>
<td>0.6</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>ACK</td>
<td>Period = 10</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Rastrigin</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Schwefel</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Ackley</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Griewank</td>
<td>Period = 20</td>
<td>0.45</td>
<td>0.35</td>
<td>0.7</td>
<td>1.5</td>
<td>0.7</td>
<td>1.5</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Schaffer</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Simulation results

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter Settings</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sphere</td>
<td>Period = 10</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.7</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Rosenbrock</td>
<td>Period = 20</td>
<td>0.85</td>
<td>0.65</td>
<td>0.7</td>
<td>1.0</td>
<td>0.9</td>
<td>0.6</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>ACK</td>
<td>Period = 10</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Rastrigin</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Schwefel</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Ackley</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>Griewank</td>
<td>Period = 20</td>
<td>0.45</td>
<td>0.35</td>
<td>0.7</td>
<td>1.5</td>
<td>0.7</td>
<td>1.5</td>
<td>0.45</td>
<td></td>
</tr>
<tr>
<td>Schaffer</td>
<td>Period = 20</td>
<td>0.95</td>
<td>0.95</td>
<td>0.7</td>
<td>1.5</td>
<td>0.5</td>
<td>1.5</td>
<td>0.95</td>
<td></td>
</tr>
</tbody>
</table>

Acknowledgments

This work was supported by JSPS Grant-in-Aid for JSPS Fellows Grant Number JP16J11745.

References


Applying Evolutionary Design of Experiments to Sensitivity Analysis of Tsunami Evacuation Simulation

Takeshi Uchitane†, Chenting Zhou‡ and Toshiharu Hatanaka‡

†Advanced Institute for Computational Science, RIKEN
7–1–26 Minatojima–minami–machi, Chuo–ku, Kobe, Hyogo 650–0047, Japan
‡Information and Physical Sciences, Osaka University
1–1 Yamadaoka, Suita, Osaka 565–0871, Japan
Email: takeshi.uchitane@riken.jp, zhou.chenting@ist.osaka-u.ac.jp, hatanaka@ist.osaka-u.ac.jp

Abstract—“Evolutionary Design of Experiments” (in short EDoE) was proposed to get an approximated result of system analysis with less number of numerical simulation executions of complex systems. In this paper, the goal is to apply the proposed method based on EDoE to tsunami evacuation simulation and evaluate the effectiveness of the proposed method. An agent based computer simulation is able to help us to make a suitable decision for actual social complex systems. In a case of evacuation from natural disasters, for instance, a planning of evacuation from tsunami in Kanazawa Japan is one of the most important political issues. Local government desires not only to estimate evacuation time but also to ensure mobility at the disaster. To evaluate the mobility, comparisons between various scenarios are required. Since actual road map contains a lot of roads, it is still difficult to check up all combinations of possible road injuries even if we can use super computers. In a context of design and analysis of experiments, selecting a set of road injuries is called design of experiments and such comparison is called sensitivity analysis. To find significant road injuries, a lot of computer resources have been required. In the proposed method, a better design which treats only significant factors is found via fitness function in terms of function optimization. To confirm the obtained designs are reasonable, numerical experiments are performed. Finally, we conclude that the proposed method enable to find better design with less experiment costs.

1. Introduction

It is important to estimate sensitivity of nonlinear systems. Especially in social systems, the number of input variables is usually large. Moreover, independences of input variables are usually unclear. Therefore, careful and comprehensive analyses of sensitivity is always required even if it takes a lot of computational cost to obtain system outputs. Evolutionary Design of Experiments[1] (in short EDoE) was proposed to get an approximated result of system analysis with less number of numerical simulation executions of complex systems. In this paper, the goal is to build a method which can be applied to tsunami evacuation simulation at Kanazawa, Japan.

An agent based computer simulation is able to help us to make a suitable decision for actual social complex systems. In a case of evacuation from natural disasters, for instance, a planning of evacuation from tsunami in Kanazawa Japan is one of the most important political issues. From an execution of agent based numerical simulation, we can obtain an estimated evacuation time in feasible elapsed time. On the other hand, local governments want to not only estimate evacuation time for various scenarios but also find significant factors which make evaluation time worse. To evaluate such significance of factors, a lot of comparisons between the scenarios are required. The number of scenarios get explosively larger as the number of system inputs increases. In the context of tsunami evacuation, availability of roads in target area is seemed as significant factors. Since actual area map contains a lot of roads, it is still difficult to check up all combinations of road injuries even if we can use super computers.

In a context of design and analysis of experiments, selecting a set of road injuries is called design of experiments and such comparison is called sensitivity analysis. As the number of system inputs gets larger, generally, more experiment costs are required to apply sensitivity analysis. Moreover, it is more difficult to distinguish significant and independent input variables. Since conventional design of experiments was proposed not for large scale models, a lot of computer resources have been required to find significant designs for such large scale models. It is because that design of experiments has been studied on wide field to apply it to systems in which independences of input variables are ensured.

In this paper, we show a case study of applying sensitivity analysis to tsunami evacuation. Then, we address a novel method to iteratively find a better design of experiments based on EDoE. Finally, better designs are found out and we conclude that the proposed method enable to find better design with less experiment costs.
2. Kanazawa Tsunami Evacuation Simulation

2.1. Summary

Tsunami evacuation time in which the target location is at Onomachi, Kanazawa, Ishikawa, Japan (See Fig. 1), was estimated[2]. At the area of Onomachi, Kanazawa, Ishikawa, Japan, earthquake–triggered tsunami may hit. Therefore, it is required that evacuation from tsunami will be finished in shorter time. However, there are many factors which may make evacuation time worse such as broken bridges and snow covered roads. If some bridges will be broken and if some road will be covered with deep snow, people will change their evacuation route. Fig. 2 shows eleven bridges which may be broken by earthquake and eleven roads which may be covered with deep snow. Evacuation time may be affected by conditions of bridges and roads. Such bridges and roads are treated as main factors. Since such bridges and roads are connected geographically, independences of main factors are not clear. Therefore, it is important to estimate significance of each main factor and significance of each interaction of main factors.

Condition of bridges are described as \( \in [0, 1] \) and condition of roads also are described as \( \in [0, 1] \). Here, 0 means that the bridge is broken or the road is covered with deep snow. Conditions of main factors are described as \([b1, b2, \ldots, b11, s1, s2, \ldots, s11]\) for eleven road adn for eleven snow road. The number of feasible combinations of conditions is \(2^{22}\) and it took about 11,650 days on single CPU core to estimate evacuation time for whole scenarios.

2.2. Sensitivity Analysis

It is generally hard to obtain \(2^{22}\) simulation results. But \(2^{22}\) simulation results can be obtained and we can apply sensitivity analysis to the simulation results. There are several method to get a result of sensitivity analysis. Here, the following linear model is employed as sensitivity analysis model.

\[
Y = XA + E
\]

Where, \( Y = [y^{(1)}, y^{(2)}, \ldots, y^{(m)}]^T \) are evacuation time, \( X=[[1, x^{(1)}_1, x^{(1)}_2, \ldots, x^{(1)}_n], \ldots, [1, x^{(m)}_1, x^{(m)}_2, \ldots, x^{(m)}_n]]^T \) are factors defined by states, \( A \) is a vector of coefficients

Table 1: Factors and their values of \( A \) which is obtained by applying regression analysis to whole simulation results are described.

<table>
<thead>
<tr>
<th>Factor</th>
<th>( A )</th>
<th>Factor</th>
<th>( A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>s7b6</td>
<td>186.9</td>
<td>s7s5</td>
<td>-23.2</td>
</tr>
<tr>
<td>s1</td>
<td>180.7</td>
<td>s4b8</td>
<td>-25.1</td>
</tr>
<tr>
<td>s4b6</td>
<td>169.1</td>
<td>s7b8</td>
<td>-25.3</td>
</tr>
<tr>
<td>b4b6</td>
<td>141.5</td>
<td>s1s5</td>
<td>-31.1</td>
</tr>
<tr>
<td>b5b6</td>
<td>141.2</td>
<td>s4s7</td>
<td>-36.2</td>
</tr>
<tr>
<td>b5b3</td>
<td>106.0</td>
<td>s4b7</td>
<td>-38.9</td>
</tr>
<tr>
<td>b4b3</td>
<td>105.7</td>
<td>b5</td>
<td>-50.3</td>
</tr>
<tr>
<td>s7b4</td>
<td>68.1</td>
<td>b4</td>
<td>-50.6</td>
</tr>
<tr>
<td>s7b5</td>
<td>67.9</td>
<td>s1b3</td>
<td>-76.7</td>
</tr>
<tr>
<td>s4b4</td>
<td>45.4</td>
<td>s4</td>
<td>-170.9</td>
</tr>
<tr>
<td>s4b5</td>
<td>45.1</td>
<td>s1b6</td>
<td>-178.8</td>
</tr>
<tr>
<td>s4b3</td>
<td>29.1</td>
<td>b3</td>
<td>-240.8</td>
</tr>
<tr>
<td>b8</td>
<td>25.2</td>
<td>s7</td>
<td>-246.4</td>
</tr>
<tr>
<td>s7b7</td>
<td>13.7</td>
<td>s1b4</td>
<td>-259.6</td>
</tr>
<tr>
<td>s4s5</td>
<td>5.4</td>
<td>s1b5</td>
<td>-259.9</td>
</tr>
<tr>
<td>s5</td>
<td>5.0</td>
<td>b5b4</td>
<td>-339.9</td>
</tr>
<tr>
<td>b7</td>
<td>1.5</td>
<td>s1s7</td>
<td>-358.6</td>
</tr>
<tr>
<td>s1b8</td>
<td>1.4</td>
<td>s6</td>
<td>-439.8</td>
</tr>
<tr>
<td>s1b7</td>
<td>-8.5</td>
<td>s1s4</td>
<td>-491.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intercept</td>
<td>4421.5</td>
</tr>
</tbody>
</table>

3. Evolutionary Design of Experiment (EDoE)

A goal of EDoE is to obtain appropriate designs for sensitivity analysis. Here, a design is described as a set of state patterns like \(0000000000000.0\) or \(00000000000000.0\). A state pattern includes several states and each state is described one among "0" or "*." Only when state is equal to "*", the associated state can be both "0" and "1". For example, with
a state pattern 00000000000000000000000000 which includes two ‘*’s, sensitivity analysis is applied to four tsunami evacuation scenarios with 000000000000000000000000000000000000 and 100000000000000000000000000000. Therefore, in sensitivity analysis, both of main factors associated with the positions of two ‘*’s and interactions of such main factors can be considered.

The best design is defined as the design which has state patterns with less number of ‘*’s and accuracy of analysis results is better. State patterns which have less number of ‘*’s require less number of simulation executions. On the other hand, since accuracy of analysis results is obtained only after applying sensitivity analysis, it is difficult to find better state patterns without background knowledges. It is because the appropriate pattern is depend on system models. Therefore, we find better designs by trial and error with state patterns evaluated by certain fitness functions. Moreover, methods to generate new designs based on the fitness function are required. The following sections, we discuss the requirements about fitness functions and methods to generate state patterns.

3.1. Fitness Function

Each state pattern must be evaluated without referring the results of other sensitivity analysis. Here, we have an assumption that a better state pattern makes variance of $Y$, (in short $V(Y)$) larger. We believe that obtaining larger values of elements of $A$ is expected when $V(Y)$ gets larger. On the other hand, simulation execution costs get higher as the number of ‘*’s included in state patterns gets larger. From the above, we consider a better state pattern makes $V(Y)$ larger and includes less number of ‘*’s.

Though the number of significant main factors is depend on system models, best design should have every significant main factors. However, without background knowledges the appropriate number of ‘*’s is unclear. Therefore, designs with several number of ‘*’s should be found. To make clear the relationships between value of $V(Y)$ and the number of ‘*’s, numerical experiments are curried out.

3.1.1. Numerical Experiments for Defining Fitness Function

Here, simplified simulation model is employed to make sure that $V(Y)$ is available to find better designs. The formulation of the simplified model is shown in equation (2) and its model parameters are shown in Table 2.

$$Y = a_0 + \sum_{i=1}^{8} a_i x_i + \sum_{i=1}^{8} \sum_{j=i+1}^{8} b_{ij} x_i x_j$$  \hspace{1cm} (2)

where, $a_0, a_1, \ldots, a_8$ are constant, $x_i$ is a model parameter and $y$ is model output. $x_i$ is input which can be equal to either 0 or 1. Values of $a_0, a_1, \ldots, a_8$ are set as the following Table 2.

<table>
<thead>
<tr>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$a_7$</th>
<th>$a_8$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1500</td>
<td>1000</td>
<td>500</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$b_1$</td>
<td>$b_2$</td>
<td>$b_3$</td>
<td>$b_4$</td>
<td>$b_5$</td>
<td>$b_6$</td>
<td>$b_7$</td>
<td>$b_8$</td>
<td></td>
</tr>
<tr>
<td>$b_1$</td>
<td>-1200</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_2$</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_3$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_4$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_5$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1000</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_6$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_7$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>$b_8$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Relationships between designs and variances are shown.

Relationships between state patterns and variances are shown in Fig. 3. From the viewpoint of obtaining better sensitivity analysis results, better state patterns which make variance of $Y$ larger are marked with a circle. Such better state patterns exist in several number of ‘*’s. In considering the simplified model, however, the best design includes four ‘*’s and its variance is largest among designs which include four ‘*’s. Therefore, we should take a strategy to find better designs which makes $V(Y)$ larger with evaluating state patterns which include less number of ‘*’s. It is because that comparison between the value of variances has less meaning in Fig. 3.

3.2. Generate New State Pattern

There is no best way to find the best design without background knowledges for the system model. But we show an strategy to find better designs. Here, a stochastic search algorithm to find better state patterns based on the strategy is proposed. Then, the algorithm is evaluated with the simplified model.

3.2.1. A Stochastic Search Algorithm

In order to find better state patterns which make their variance larger, methods like Genetic Algorithm’s crossover and mutation are applied in generating new state
patterns. Pseudo-code of proposed algorithm is described as follows.

1. generate initial state patterns which have just 2 "*"s $(t = 0)$
2. evaluate variance of $Y$ for each state pattern
3. generate new state pattern which must include $\text{NA}(t)$ or less number of "*"s and must not same as previously generated state patterns
4. if terminated condition is satisfied then exit, else return 2

3.2.2. Numerical Experiments for Evaluating Proposed Algorithm

A new state pattern is generated via the following processes. In crossover process, a pair of state patterns is selected by using rank 2 tournament selection among survived state patterns and each state in a new state pattern is ether state of selected one or another stochastically. In mutation process, for each state of the new state pattern, the value is replaced with "*" with certain mutation rate. Since crossover and mutation processes may increase number of "*"s, "*" in the new state pattern is randomly replaced with 0 until the number of "*"s is less or equal to $\text{NA}(t)$. Here, $\text{NA}(t)$ is a function that returns an integer value and the returned value increase as $t$ gets large. When maximum value of $t$ is 10, $\text{NA}(t \leq 2) = 3$ and $\text{NA}(t \leq 10) = 4$ are defined. New state patterns are generated until the number of unique state patterns is same as the number of initial state patterns. Variances of generated designs are evaluated and designs which makes $V(Y)$ less are died until the number of existing designs is more than the number of initial designs.

Fig. 4 shows the result of applying proposed methods. Best state patterns which include three or less number of "*"s were found in early generation. On the other hand, the best state pattern which includes four "*"s was found in 9th generation. Since the best state pattern was found, this search process as a design is good.

4. Discussion

For a simplified model, proposed algorithm could find a better design without treating whole feasible designs. In order to apply evolutionary design of experiment, fitness function and methods to generate state patterns are required. Definition of fitness function and definition of the methods may affect the search process as design. Moreover, the best state pattern was found in later generation. Therefore, effectiveness of algorithm to find better designs in early generation is significant.

5. Conclusion

Evolutionary Design of Experiments is a novel method to find appropriate designs without evaluating whole feasible input values. A simplified model was built in order to evaluate the proposed algorithm. For a simplified model, it seems that proposed algorithm could find a better design with less number of model evaluations.

The proposed algorithm can be applied to make a design for Kanazawa tsunami evacuation problem. Applied algorithm and evaluate obtained designs will be shown in presentation.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP15K17502. Tsunami evacuation simulation results were provided by collaborators of the other study who are T. Yamashita, J. Tsuji, H. Matsushima, I. Noda and N. Ito.

References

Hybrid Method of Genetic Algorithm and Firefly Algorithm Distinguishing between Males and Females

Masaki Takeuchi†, Haruna Matsushita‡, Yoko Uwate† and Yoshifumi Nishio†

†Dept. of Electrical and Electronic Engineering, Tokushima University
2-1 Minami-Josanjima, Tokushima 770–8506, Japan
‡Dept. of Electronics and Information Engineering, Kagawa University
2217-20 Hayashi-cho, Takamatsu, Kagawa, 761-0396, Japan
Email: masaki@ee.tokushima-u.ac.jp, haruna@eng.kagawa-u.ac.jp, {uwate, nishio}@ee.tokushima-u.ac.jp

Abstract—We have proposed Firefly Algorithm Distinguishing between Males and Females (FADMF). This algorithm exists together with males and females. In this study, we propose Firefly Algorithm Distinguishing between Males and Females Combined with Genetic Algorithm (FADCG). This proposed algorithm is applied genetic operators every certain iteration. We compare these two algorithms and the conventional Firefly Algorithm by using 2013 Congress on Evolutionary Computation (CEC) benchmark functions. Numerical experiments indicate that FADCG is effective for complex optimization problems.

1. Introduction

Evolutionary Computing (EC) is a subfield of artificial intelligence (AI) in computer science, and is based on biological mechanisms of evolution. EC technique mainly involves metaheuristic optimization algorithms such as Evolutionary Algorithm (EA) and Swarm Intelligence (SI).

Genetic Algorithm (GA) is one paradigm and most popular technique of major EA. On GA, individuals of a population evolve according to crossover, mutation, and selection from the population. The crossover and mutation create the necessary diversity. On the other hand, selection acts as a force increasing quality. There are two most notable advantage: the ability of dealing with complex problems and parallelism. However, GA also has some minor disadvantages. The choice of important parameters such as the mutation probability and the crossover probability, and the selection criteria of new population should be carefully carried out.

SI algorithm is one of stochastic algorithms. Stochastic algorithms have a deterministic component and a random component. Algorithms having only the deterministic component are almost all local search algorithms. There is a risk to be trapped at local optima such algorithms. However, stochastic algorithms are possible to jump out such locality. SI algorithms are based on the idealized behavior of animals and insects. Representative examples are Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and Firefly Algorithm (FA) [1–3].

On FA, all fireflies are unisex. However, there are males and females in the real world. In our previous study, we have proposed algorithm distinguishing sex of fireflies [4]. This method is called Firefly Algorithm Distinguishing between Males and Females (FADMF). On FADMF, the movements of males and females are different from each other. This proposed algorithm has been applied 27 benchmark functions of Congress on Evolutionary Computation (CEC) 2013. Numerical experiments indicate that FADMF is superior to the conventional FA under some conditions. FADMF jump out locality more easily than the conventional FA, while FA-DMF is inferior about absorption speed.

It is paid many attentions to combine SI algorithm with GA [5, 6]. These hybrid algorithms outperform the standard algorithms. Especially, some hybrid PSO and GA algorithms obtain better results than the conventional PSO and GA. In this study, we propose the hybrid method of FADMF and GA. This method is called Firefly Algorithm Distinguishing between Males and Females Combined with Genetic Algorithm (FADCG). This method involves male and female swarms, and is also performed every certain iteration by using genetic operators. We compare the proposed method and the conventional FA by using 27 benchmark functions of CEC 2013. Numerical experiments indicate that the proposed method is more efficient algorithm than the conventional FA.

This study is organized as follows: first, we explain the conventional Genetic Algorithm in Section 2, and then, we explain the conventional Firefly Algorithm in Section 3. The next, we explain FADMF in Section 4. Followed by, we describe in detail of FADCG. In Section 6, we show numerical experiments. Finally, we conclude in this study.

2. Firefly Algorithm (FA)

Firefly Algorithm (FA) has been developed by Yang, and it was based on the idealized behavior of the flashing characteristics of fireflies. The conventional FA is idealized these flashing characteristics as the following three rules

• All fireflies are unisex so that one firefly is attracted to other fireflies regardless of their sex;
Attractiveness is proportional to their brightness, thus for any two flashing fireflies, the less brighter one will move towards the brighter one. The attractiveness is proportional to the brightness and they both decrease as their distance increases. If no one is brighter than a particular firefly, it moves randomly.

The brightness or light intensity of a firefly is affected or determined by the landscape of the objective function to be optimized.

Attractiveness of firefly \( \beta \) is defined by

\[
\beta = \beta_0 e^{-\gamma r_{ij}}
\]

where \( \gamma \) is the light absorption coefficient, \( \beta_0 \) is the attractiveness at \( r_{ij} = 0 \), and \( r_{ij} \) is the distance between any two fireflies \( i \) and \( j \) at \( x_i \) and \( x_j \). The movement of the firefly \( i \) is attracted to another more attractive firefly \( j \), and is determined by

\[
x_i = x_i + \Delta x, \Delta x = \beta(x_j - x_i) + \alpha \epsilon_i,
\]

where \( x_i \) is the position vector of firefly \( i \), \( \epsilon_i \) is the vector of random variable, and \( \alpha(t) \) is the randomization parameter. The parameter \( \alpha(t) \) is defined by

\[
\alpha(t) = \alpha(0) \left( \frac{10^{-4}}{0.9} \right)^{t/\text{max}},
\]

where \( t \) is the number of iteration.

3. Genetic Algorithm (GA)

Genetic Algorithm (GA) has been developed by Holland J. and it is a model of biological evolution based on Charles Darwin’s theory of nature selection. The conventional GA is often done by the following procedure:

- Creating a population of individuals;
- Evaluating the fitness of all the individuals in the population;
- Updating the population;
- Terminating generation when a maximum number of generations has been produced.

The essential part of GA is formed from genetic operators such as the crossover, mutation, and selection.

Individuals are stochastically selected from the population to create the basis of the next population. The fitter individuals have a more chance of selection than weaker one. There are many ways how to select the best individuals, such as Roulette Wheel Selection, Rank Selection, and Tournament Selection. The crossover selects genes from parent, and creates a new offspring. Commonly, a process of taking two parent genes is used, such as two-point crossover and uniform crossover. After the crossover is performed, individuals are mutated. This process is to prevent falling into a local optima. Genes of the offspring are changed randomly by the mutation.

The crossover probability, the mutation probability, and population size should be carefully carried out.

4. Firefly Algorithm Distinguishing between Males and Females (FADMF)

One of the rules of the conventional FA is all fireflies are unisex. However, males and females exist in the real world. Therefore, we distinguish sex of fireflies, that is, there are two swarms in our proposed method. We call our proposed method Firefly Algorithm Distinguishing between Males and Females (FA-DMF). The movement of female is modeled from the physical differences. In the real world, females are bigger than males and female eyes are smaller than male. Thus, in our proposed method, females move slower than males, and females have difficulty finding the flashes of other distant fireflies. In addition, we change the randomization parameter of female.

The female parameters \( \alpha(t) \) and \( \beta \), and the female movement \( x \) is determined with parameters \( V \) and \( W \) by

\[
\alpha(t) = \alpha(0) \left( \frac{10^{-4}}{0.9} \right)^{t/\text{max}},
\]

\[
\beta = \beta_0 e^{-\gamma r_{ij}/W},
\]

\[
x = x + \Delta x/V.
\]

In the proposed method, males are attracted to all fireflies, while females are attracted to only males. Males move the same as fireflies of the conventional FA.

5. Firefly Algorithm Distinguishing between Males and Females Combined with Genetic Algorithm (FADCG)

In this study, we propose the hybrid method of FADMF and GA. This proposed method is called Firefly Algorithm Distinguishing between Males and Females Combined with Genetic Algorithm (FADCG). All fireflies move every iteration according to FADMF. In addition, fireflies are applied genetic operators every certain iteration. We use uniform crossover and the mutation of genetic operators. We define that the crossover probability is 100 percent and the mutation probability is 30 percent.

6. Numerical Experiments

We compare FADCG to the conventional FA and FADMF with benchmark functions of CEC 2013 except function 20 (see Table 1).
Table 1: Benchmark Functions of CEC 2013

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>( f(x^*) )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Unimodal Functions</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Sphere function</td>
<td>-1400</td>
</tr>
<tr>
<td>2</td>
<td>Rotated High Conditioned Elliptic Function</td>
<td>-1300</td>
</tr>
<tr>
<td>3</td>
<td>Rotated Bent Cigar Function</td>
<td>-1200</td>
</tr>
<tr>
<td>4</td>
<td>Rotated Discus Function</td>
<td>-1100</td>
</tr>
<tr>
<td>5</td>
<td>Different Powers Function</td>
<td>-1000</td>
</tr>
<tr>
<td></td>
<td>Basic Multimodal Functions</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Rotated Rosenbrock’s Function</td>
<td>-900</td>
</tr>
<tr>
<td>7</td>
<td>Rotated Schaffers F7 Function</td>
<td>-800</td>
</tr>
<tr>
<td>8</td>
<td>Rotated Ackley’s Function</td>
<td>-700</td>
</tr>
<tr>
<td>9</td>
<td>Rotated Weierstrass Function</td>
<td>-600</td>
</tr>
<tr>
<td>10</td>
<td>Rotated Griewank’s Function</td>
<td>-500</td>
</tr>
<tr>
<td>11</td>
<td>Rastrigin’s Function</td>
<td>-400</td>
</tr>
<tr>
<td>12</td>
<td>Rotated Rastrigin’s Function</td>
<td>-300</td>
</tr>
<tr>
<td>13</td>
<td>Non-Continuous Rotated Rastrigin’s Function</td>
<td>-200</td>
</tr>
<tr>
<td>14</td>
<td>Schwefel’s Function</td>
<td>-100</td>
</tr>
<tr>
<td>15</td>
<td>Rotated Schwefel’s Function</td>
<td>100</td>
</tr>
<tr>
<td>16</td>
<td>Rotated Katsuura Function</td>
<td>200</td>
</tr>
<tr>
<td>17</td>
<td>Lunacek Bi Rastrigin Function</td>
<td>300</td>
</tr>
<tr>
<td>18</td>
<td>Rotated Lunacek Bi Rastrigin Function</td>
<td>400</td>
</tr>
<tr>
<td>19</td>
<td>Expanded Griewank’s plus Rosenbrock’s Function</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td>Composition Functions</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>Composition Function 1 (n=5, Rotated)</td>
<td>700</td>
</tr>
<tr>
<td>22</td>
<td>Composition Function 2 (n=3, Unrotated)</td>
<td>800</td>
</tr>
<tr>
<td>23</td>
<td>Composition Function 3 (n=3, Rotated)</td>
<td>900</td>
</tr>
<tr>
<td>24</td>
<td>Composition Function 4 (n=3, Rotated)</td>
<td>1000</td>
</tr>
<tr>
<td>25</td>
<td>Composition Function 5 (n=3, Rotated)</td>
<td>1100</td>
</tr>
<tr>
<td>26</td>
<td>Composition Function 6 (n=5, Rotated)</td>
<td>1200</td>
</tr>
<tr>
<td>27</td>
<td>Composition Function 7 (n=5, Rotated)</td>
<td>1300</td>
</tr>
<tr>
<td>28</td>
<td>Composition Function 8 (n=5, Rotated)</td>
<td>1400</td>
</tr>
</tbody>
</table>

The optimal solutions \( x^* \) of these benchmark functions is shifted from 0, and the global optima \( f(x^*) \) are not equal to 0. The search range of these functions is \([-100, 100]^D\), and the dimension \( N \) is 30. Each numerical experiment is run 50 times. In each test function, \( l_{max} = 1500 \), \( V = 3 \), \( W = 4 \). In this study, we change female percentage from 10 to 90 every 10 percentage. The best female percentage of FADCG is 40 percent, while the best female percentage of FADMF is 30 percent. Numerical experiments of the best female percentage are summarized in Table 2. Table 2 shows the average value, minimum value, maximum value, and standard deviation.

FADCG obtains a lot of best solutions more than other two algorithms. FADCG performs best on 10 times. In addition, FADCG obtains better results than the conventional FA at 17 times. Therefore, FADCG is superior to the conventional FA and FADMF.

In the case of unimodal functions, FADMF performs best on 3 times. FADCG obtains significantly worse results than other two algorithms. Therefore, we assume that FADCG converges slower than the conventional FA and FADMF.

In the case of basic multimodal functions, FADMF and FADCG perform best on 5 times. Therefore, FADMF and FADCG are fitted for basic multimodal functions.

In the case of composition functions, FADCG performs best on 5 times. FADCG is significantly superior to other two algorithms.

7. Conclusion

In this study, we have proposed Firefly Algorithm Distinguishing between Males and Females Combined with Genetic Algorithm (FADCG). This algorithm has male and female swarms which move differently each other, and is applied genetic operators every certain iteration. We have compared FADCG to the conventional FA and FADMF by using benchmark functions of Congress on Evolutionary Computation (CEC) 2013. Numerical experiments indicate that FADCG is superior to other algorithms, while FADMF is superior to other functions for unimodal functions. In other words, FADCG is effective for complex multimodal functions.

In the future work, we investigate parameters of FADCG more details. Furthermore, we compare FADCG to other improved algorithms, and apply to actual optimization problems.

References


f
f1

f2

f3

f4

f5

f6

f7

f8

f9

f10

f11

f12

f13

f14

avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std
avg
min
max
std

FA
6.11 × 10−4
3.47 × 10−4
9.53 × 10−4
1.40 × 10−4
1.09 × 107
5.31 × 106
1.93 × 107
3.39 × 106
1.37 × 107
3.77 × 103
1.74 × 108
2.54 × 107
1.09 × 105
6.22 × 104
1.57 × 105
2.54 × 104
3.89 × 101
2.28 × 10−2
1.53 × 102
3.19 × 101
2.74 × 101
2.60 × 101
2.92 × 101
7.23 × 10−1
1.16 × 101
2.39 × 100
3.01 × 101
6.14 × 100
2.14 × 101
2.12 × 101
2.16 × 101
7.45 × 10−2
1.00 × 101
4.86 × 100
1.48 × 101
2.43 × 100
6.29 × 10−1
7.31 × 10−2
2.20 × 100
5.41 × 10−1
2.68 × 101
1.29 × 101
5.07 × 101
7.12 × 100
3.01 × 101
1.49 × 101
5.17 × 101
8.16 × 100
6.77 × 101
1.52 × 101
1.18 × 102
2.52 × 101
2.34 × 103
9.36 × 102
4.15 × 103
6.73 × 102

FADMF
5.61 × 10−4
2.34 × 10−4
9.33 × 10−4
1.25 × 10−4
3.96 × 106
1.39 × 106
7.96 × 106
1.64 × 106
9.14 × 106
5.18 × 103
5.08 × 107
1.25 × 107
1.38 × 105
8.09 × 104
2.09 × 105
2.87 × 104
1.92 × 10−2
1.36 × 10−2
2.58 × 10−2
3.46 × 10−3
2.65 × 101
2.05 × 101
7.72 × 101
7.49 × 100
4.96 × 100
4.30 × 10−1
2.03 × 101
5.04 × 100
2.14 × 101
2.12 × 101
2.16 × 101
1.03 × 10−1
1.01 × 101
4.11 × 100
1.62 × 101
2.42 × 100
1.95 × 10−1
1.58 × 10−2
1.03 × 100
2.20 × 10−1
3.88 × 101
1.59 × 101
6.07 × 101
1.07 × 101
3.77 × 101
1.49 × 101
6.17 × 101
9.42 × 100
9.43 × 101
3.96 × 101
1.53 × 102
2.83 × 101
2.20 × 102
1.26 × 103
3.23 × 103
4.08 × 102

Table 2: Numerical Experiments
FADCG
avg
2.26 × 103
1.06 × 10−3
−4
min
1.20 × 103
5.70 × 10
f15
−3
max
3.80 × 103
1.64 × 10
−4
std
5.63 × 102
2.44 × 10
6
avg
9.28 × 10−2
6.59 × 10
6
min
3.17 × 10−2
1.94 × 10
f16
7
max
2.28 × 10−1
1.50 × 10
6
std
3.99 × 10−2
2.97 × 10
7
avg
5.95 × 101
1.19 × 10
4
min
4.79 × 101
2.34 × 10
f17
7
max
7.52 × 101
9.01 × 10
7
std
7.10 × 100
1.94 × 10
5
avg
6.27 × 101
2.57 × 10
5
min
4.85 × 101
1.37 × 10
f18
5
max
8.62 × 101
4.05 × 10
4
std
8.16 × 100
6.07 × 10
−1
avg
3.77 × 100
1.78 × 10
−2
min
2.44 × 100
6.52 × 10
f19
−1
max
6.13 × 100
2.87 × 10
−2
std
8.19 × 10−1
5.30 × 10
1
avg
3.30 × 102
2.73 × 10
1
min
2.00 × 102
2.60 × 10
f21
1
max
4.44 × 102
2.94 × 10
−1
std
8.52 × 101
8.15 × 10
avg
3.31 × 103
6.35 × 100
min
1.32 × 103
8.51 × 10−1
f
22
max
6.17 × 103
1.72 × 101
std
1.14 × 103
4.02 × 100
avg
3.84 × 103
2.14 × 101
min
2.40 × 103
2.13 × 101
f
23
max
5.75 × 103
2.15 × 101
std
8.43 × 102
7.11 × 10−2
avg
2.17 × 102
8.74 × 100
min
2.01 × 102
3.25 × 100
f
24
max
2.41 × 102
1.45 × 101
std
1.16 × 101
2.40 × 100
avg
2.34 × 102
5.50 × 10−1
min
2.20 × 102
6.22 × 10−2
f
25
max
2.51 × 102
2.13 × 100
std
7.73 × 100
4.99 × 10−1
avg
2.89 × 102
2.42 × 101
min
2.00 × 102
15.0 × 101
f
26
max
3.34 × 102
3.78 × 101
std
4.78 × 101
4.92 × 100
avg
4.56 × 102
2.96 × 101
1
min
3.13 × 102
1.49 × 10
f27
1
max
6.59 × 102
5.17 × 10
0
std
1.17 × 102
7.95 × 10
avg
3.06 × 102
7.85 × 101
1
min
1.01 × 102
2.29 × 10
f28
2
max
1.36 × 103
1.36 × 10
1
std
1.60 × 102
2.45 × 10
3
f
FA
1.51 × 10
6.91 × 102
best solution
8
2.75 × 103
more than the conventional FA
4.21 × 102

- 545 -

2.23 × 103
1.22 × 103
3.44 × 103
4.76 × 102
1.29 × 10−1
4.94 × 10−2
2.24 × 10−1
4.65 × 10−2
8.71 × 101
7.04 × 101
1.31 × 102
1.42 × 101
9.16 × 101
6.69 × 101
1.27 × 102
1.47 × 101
4.01 × 100
2.45 × 100
6.20 × 100
9.78 × 10−1
3.39 × 102
1.01 × 102
4.44 × 102
9.12 × 101
2.61 × 103
7.27 × 102
4.51 × 103
7.70 × 102
3.31 × 103
1.37 × 103
5.69 × 103
9.94 × 102
2.22 × 102
2.01 × 102
2.39 × 102
9.49 × 100
2.32 × 102
2.01 × 102
2.53 × 102
1.23 × 101
2.85 × 102
2.00 × 102
3.35 × 102
5.34 × 101
5.17 × 102
3.14 × 102
7.28 × 102
9.32 × 101
3.09 × 102
1.00 × 102
1.32 × 103
2.17 × 102
FADMF
9
13

2.52 × 103
1.33 × 103
4.45 × 103
5.65 × 102
3.30 × 10−1
9.30 × 10−2
7.01 × 10−1
1.35 × 10−1
7.59 × 101
5.88 × 101
1.12 × 102
1.14 × 101
8.61 × 101
5.99 × 101
1.36 × 102
1.50 × 101
3.47 × 100
2.03 × 100
4.91 × 100
6.67 × 10−1
3.12 × 102
2.00 × 102
4.44 × 102
8.82 × 101
1.71 × 103
8.51 × 102
2.34 × 103
4.02 × 102
2.99 × 103
1.22 × 103
4.74 × 103
7.00 × 102
2.23 × 102
2.03 × 102
2.34 × 102
7.52 × 100
2.25 × 102
2.14 × 102
2.40 × 102
5.47 × 100
3.00 × 102
2.00 × 102
3.34 × 102
4.39 × 101
4.85 × 102
3.26 × 102
6.23 × 102
9.08 × 101
2.97 × 102
1.01 × 102
3.02 × 102
2.81 × 101
FADCG
10
17


Tabu Search Method
for Solving Covering Salesman Problem with Nodes and Segments

Takafumi Matsuura
Faculty of Engineering, Nippon Institute of Technology
4–1 Gakuendai, Miyashiro-machi, Minamisaitama-gun, Saitama, 345-8501 JAPAN
E-mail: matsuura@nit.ac.jp

Abstract—As a new covering problem, we have already mathematically formulated Covering Salesman Problem with Nodes and Segments (CSPNS). In the CSPNS, a city distribution is given. The goal of the CSPNS is to identify the shortest tour of a subset of all given cities, such that each city which is not on the tour is within radius $r$ of any city or segment on the tour. In addition, we have already proposed a local search method for the CSPNS. In this paper, to find better solutions of the CSPNS, we propose a heuristic method by using a tabu search.

1. Introduction

The Covering Salesman Problem with Nodes and Segments (CSPNS) is a new covering problem [1]. In the CSPNS, a city distribution is given. The CSPNS is to identify the shortest tour of a subset of all given cities, such that a city which is not on the tour is within radius $r$ of any city (node) or path (segment) on the tour. To solve the CSPNS, we have already proposed a local search method [1]. In the method, first, an initial tour passing through all cities is constructed, that is, the Traveling Salesman Problem (TSP) is solved. Second, many visited cities are eliminated from the tour. Finally, a visited city and an unvisited city are exchanged to make a shorter tour. Although, the method quickly finds solutions of the CSPNS, the obtained solutions are local optimal solutions.

To find better solutions of combinatorial optimization problems, many metaheuristics were proposed such as the simulated annealing (SA) [2], the genetic algorithm (GA) [3], and the tabu search (TS) [4–6]. Among them, the tabu search shows good performances for many combinatorial optimization problems for example traveling salesman problem [6, 7], quadratic assignment problem [8, 9], and vehicle routing problem [10]. In this paper, to find better solutions of the CSPNS, we propose a new method by using the tabu search. As a result, the proposed method shows good performances for the CSPNS.

2. Covering Salesman Problem with Nodes and Segments

In the CSPNS, a set of cities $V = \{1, 2, \ldots, n\}$, distances $d_{ij}$ between city $i$ and city $j$, perpendicular distances $c_{ijk}$ between city $i$ and path $jk$, and a covering distance $r > 0$ are given. If distance $d_{ij}$ is less than or equal to $r$, a city $i$ can cover a city $j$. If perpendicular distance $c_{ijk}$ is less than or equal to $r$, a path $jk$ can cover a city $i$. The constraint conditions of the CSPNS are as follows: (1) the salesman starts from the city 1 and goes back to the city 1, (2) the salesman can visit each city at most once, and (3) all cities are within radius $r$ of any visited city or path on the tour. The goal of the CSPNS is to identify the shortest tour which satisfies the constant conditions. Figure 1 shows a graphical example of the CSPNS.

3. Proposed Methods

3.1. Local Search Methods

We have already proposed local search methods [1] for solving the CSPNS. In the method, first, an initial tour passing through all given cities is constructed, that is, the TSP is solved. Second, the length of the initial tour is improved by a local search method. Next, many visited cities are eliminated from the tour. Finally, a visited city and an unvisited city are exchanged to make a shorter tour.

The procedure of the local search method is shown as follows [1]:

Step 1: Constructing an initial tour
A random tour passing through all given cities is constructed.

**Step 2: Improving the length the initial tour**

To make a shorter tour, the initial tour is improved by the Lin-Kernighan heuristic [11]. The Lin-Kernighan heuristic is applied until no further improvement cannot be obtained.

**Step 3: Removing algorithm**

A visited city \(i\) is removed from a current tour, if a new tour is feasible solutions (Fig.2(a)). If many good improvements can be found, the best improvement is carried out.

**Step 4: Exchanging algorithm**

To make a shorter tour, a visited city \(i\) and an unvisited city \(j\) are swapped (Fig.2(b)). If many good combinations are found, the best exchange is carried out.

**Step 5: Swapping algorithm**

To make a shorter tour, a visited city \(i\) and other visited city \(j\) are swapped (Fig.2(c)).

**Step 6: Repeating each local search**

Until a local optimum solution is obtained, the removing algorithm, the exchanging algorithm, and the swapping algorithm are repeated.

3.2. A Method by using Tabu Search

To find good near-optimal solutions of the CSPNS, we propose a metaheuristics method by using the tabu search [4–6]. The local search method moves from one solution \(S\) to another improving solution \(S'\) in neighborhood solutions of the solution \(S\), until a local optimum solution is found. To escape from the local optimum solution, the tabu search moves from one solution \(S\) to the best solution \(S'\) in neighborhood solutions of the solutions \(S\), even though the solution \(S'\) is worse than the solution \(S\). Then, to avoid period exploration process, the solution \(S\) is added to a tabu list. The information of the tabu list is used to guide the move from a current solution to the next solution. The solutions in the tabu list cannot be selected as a next solution for a certain temporal duration. Its duration is called tabu tenure.

The proposed method has four movements: (i) removing step, in which a visited city is removed from the tour (Fig.2(a)), (ii) exchanging step, in which a visited city in the tour is exchanged by an unvisited city (Fig.2(b)), (iii) swapping step, in which one visited city and another visited city are swapped (Fig.2(c)), and (iv) adding step, in which an unvisited city is added into the tour (Fig.2(d)). In the CSPNS, if an unvisited city is added into a current tour, the length of the new tour always becomes longer. However, when the number of visited cities is low, it is difficult to
construct feasible neighborhood solutions in the removing step, the swapping step, and the exchanging step. Therefore, we add the unvisited city into the tour at the adding step.

Then, three tabu lists are constructed in the method. One tabu list memories a deleted city from the tour (Deleted list). The cities listed in Deleted list cannot be added to the tour for \( \tau_d \) tenure. Another list memories an added city into the tour (Added list). The cities listed in Added list cannot be deleted from the tour for \( \tau_a \) tenure. The other list memories swapped cities in the swapped step (Swapped list). The cities listed in Swapped list cannot be exchanged for \( \tau_s \) tenure in Exchanging step. In the removing step, when a visited city \( i \) is removed from the tour, the city \( i \) is listed in the Deleted list. In the exchanging step, when a visited city \( i \) is exchanged for an unvisited city \( j \), the city \( i \) is listed in the Deleted list and the city \( j \) is listed in the Added list. In the swapping step, when visited cities \( i \) and \( j \) are swapped, the city \( i \) is listed in the Swapped list. In the proposed method, the removing step, the exchanging step, the swapping step and the adding step are carried out just one time in one iteration. In the proposed method, if the best solution is updated, the corresponding tour is improved to obtain a local optimal solution by the proposed local search methods.

4. Simulations and Results

To investigate performances of the proposed method, we used DIMACS [12] which is one of the benchmark problems for the Traveling Salesman Problem. The number of city \( n \) is set to 50 and 100. The cities are uniformly distributed in the \( 10^6 \times 10^6 \) square and seed for making the instances is set to 1.

The covering distance \( r \) is set to 20,000, 40,000, 60,000, and 80,000. The tabu tenure \( \tau_d \), \( \tau_a \), and \( \tau_s \) are set to several values. In this simulation, the values of \( \tau_a \) and \( \tau_s \) are set to same values. The iteration of the tabu search method is 500. We conducted simulations using the Intel compiler on a Mac Pro (2.8 GHz Intel Core i7) with 16GB memory running Mac OS X 10.10.5.

By using the formulation of the CSPNS [1] and a mixed-integer programming solver, we can obtain an optimal solution. Table 1 shows a length of an optimal tour or a near-optimal tour obtained by the gurobi optimizer (Table 1) in one hour. From Fig.3, when the length of tabu tenure is small, the proposed method cannot obtain good solutions because the method cannot escape local minima. When we appropriately set to the length of the tabu tenure, the proposed method obtained good results for all instances and the covering distance.

<table>
<thead>
<tr>
<th>( n )</th>
<th>Covering distance ( r )</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>5,973,311</td>
</tr>
<tr>
<td>100</td>
<td>7,865,796</td>
</tr>
</tbody>
</table>

Table 2 shows the attained results for the local search method and the tabu search method. Results are expressed by percentages of average gaps from obtained solution and the best solution obtained by the gurobi optimizer (Table 1).

<table>
<thead>
<tr>
<th>( r )</th>
<th>Local Search</th>
<th>Tabu Search</th>
<th>Improvement rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>20,000</td>
<td>2.83</td>
<td>2.01 (16,14)</td>
<td>0.95 %</td>
</tr>
<tr>
<td>40,000</td>
<td>3.41</td>
<td>2.01 (12,14)</td>
<td>1.35 %</td>
</tr>
<tr>
<td>60,000</td>
<td>2.76</td>
<td>1.53 (9,15)</td>
<td>1.19 %</td>
</tr>
<tr>
<td>80,000</td>
<td>9.43</td>
<td>5.57 (2,16)</td>
<td>3.52 %</td>
</tr>
</tbody>
</table>

5. Conclusions

In this paper, we proposed a metaheuristics method by using the tabu search for solving the CSPNS. From the computational results, although the proposed method uses simple local search methods and the tabu search, it obtains good solutions. In the future work, it is important to develop an effective adjustment method of tabu tenure. We
also develop variable neighborhood search to find good solutions. It is desirable to solve much larger problems, such as $10^6$ order problems.

References


Figure 3: Performance of the proposed method. The average gaps (%) between an optimal solution (a near-optimal solution) obtained by the gurobi optimizer and solutions of the proposed method in 50 trials are indicated by color bars.
Tracking optima in dynamical problems
by a chaotic population based optimizer

Yoshikazu YAMANAKA† and Tadashi TSUBONE‡

†‡Nagaoka University of Technology, 1603-1, Kamitomioka, Nagaoka, Niigata, 940-2188, Japan
†Email: y.yama@stn.nagaokaut.ac.jp, ‡Email: tsubone@vos.nagaokaut.ac.jp

Abstract—Tracking the optimal solution is required in dynamic optimization problems. Some existing methods can track the optimum by detection methods based on re-evaluation, however, some issues of the detection are reported. We have proposed tolerance memory update and applied it to our previous proposed optimizer. Without the detection, the propose method achieves the tracking.

1. Introduction

Some of the optimization problems are dynamical whose object functions have multiple peaks and the peaks can be changed. To solve these problems, algorithms are required two important abilities. One of them is tracking the bottom or the top of the single peak. Another is searching ability for another peak whose bottom or top denotes better fitness than current tracking peaks one. In this paper, we focus on the first ability because the second requirement can be obtained by expansion, such as using multi-swarm, clustering and hybridizing with other searching mechanisms, based on the methods that have the first ability [1].

Focusing on the first ability, multi-population based optimizer, such as particle swarm optimization (PSO) [2, 3] and downhill simplex (DS) [4], have following issues to be solved [5]. These methods keep fitness of the best position in searching history, however, it may no longer be true when the function changes. This is called outdated memory. The swarm and the simplex converge to a specific point with sufficiently iterations, then, will inhibit tracking the shifted optimum. This is diversity loss issue. To solve them, detection methods by re-evaluating the best position is proposed, however, the performance depends on the detection time and decreases in functions with noise [1].

In this paper, we have proposed tolerance memory update to solve the outdated memory, and applied it to an optimizer based on piecewise-rotational chaotic system (OPRC) [6], which has no diversity loss issue because of the chaotic behavior. OPRC with the tolerance memory update achieves tracking without the detection, and it obtains better performance than compared methods.

2. Optimization methods

In this section, PSO, atomic PSO (APSO) [7], OPRC, and DS are introduced. These algorithms update some searching points with time-step \(t\), that are candidates of the optimal solution. These methods are originally proposed for static problems. Here, we consider these algorithms for dynamical problems that can be changed by number of evaluation times \(n_e\). Let \(f(x_i(t), n_e)\) be a \(d\)-dimensional dynamical object function, where \(x_i(t) \in \mathbb{R}^d\) is a searching point, and \(i = 1, 2, \ldots, n\) is index of the searching points. In this paper, let \(f(x_i(t), n_e)\) be a maximizing problem.

2.1. Particle swarm optimization (PSO)

PSO searches the optimal solution by updating \(x_i(t)\) and an independent variable, \(v_i(t)\). The fitness of \(x_i(t)\) are stored by \(n\)-dimensional vector \(f_x\). Let \(p_b\) be \(x_i(t)\) that denotes the best fitness value in the \(i\)-th searching points history, and let \(g_b\) be \(p_b\) that denotes the best fitness in all searching points history. The fitness of \(p_b\) and \(g_b\) are stored by a \(n\)-dimensional vector \(f_p\) and a scalar \(f_g\), respectively. \(x_i(t)\) and \(v_i(t)\) are updated by following dynamical system:

\[
\begin{align*}
    v_{ij}(t + 1) &= \omega v_{ij}(t) + c_1 r_{i1}[p_{bij} - x_{ij}(t)] + c_2 r_{i2}[g_{bij} - x_{ij}(t)] \quad (1) \\
    x_{ij}(t + 1) &= v_{ij}(t + 1) + x_{ij}(t),
\end{align*}
\]

where \(j\) is index of the vector, \(\omega, c_1\) and \(c_2\) are system parameters, \(r_{i1}\) and \(r_{i2}\) are random values with uniform distribution \([0, 1]\). In this paper, \(\omega = 0.729, c_1 = c_2 = 1.49445\) [8] are selected. Using these parameters, the dynamical system seems to be weekly stable. The algorithm is described in Algorithm 1.

2.2. Atomic PSO (APSO)

APSO is a modified algorithm based on PSO by adding an acceleration vector into the dynamical system in order to prevent the diversity loss [7]. An acceleration vector of \(i\)-th searching point \(a_i \in \mathbb{R}^d\) is given by

\[
a_i = \begin{cases}
    \sum_{k 
eq i} \frac{q_i q_k}{r_k} (x_k - x_i) & \text{for } p_{\text{min}} < r_k < p_{\text{max}}, \\
    0 & \text{otherwise},
\end{cases}
\]

(3)

where \(q_i\) is the charge parameter for \(i\)-th searching point, \(p_{\text{min}}\) and \(p_{\text{max}} = (x_{\text{max}})^{1/3}\) are system parameters, and \(r_k = \|x_k - x_i\|\), where \(\|\cdot\|\) is norm. \(v_{ij}(t)\) of APSO is updated by

\[
\begin{align*}
    v_{ij}(t + 1) &= \omega v_{ij}(t) + c_1 r_{i1}[p_{bij} - x_{ij}(t)] + c_2 r_{i2}[g_{bij} - x_{ij}(t)] + a_{ij},
\end{align*}
\]

(4)
The algorithm of APSO is same as Algorithm 1 except procedure 2: \((x_i(t), v_i(t)) \rightarrow (x_i(t+1), v_i(t+1))\) by Eqs. (4 and 2). In this paper, \(p_{min} = 1, v_{max} = 100, q_i = 16\) for 50\% of the searching points, and \(q_i = 0\) for the others [7]. Because of \(a_i\), the diversity loss can be prevented.

2.3. An optimizer based on piecewise-rotational chaotic system (OPRC)

OPRC updates \(x_i(t)\) and \(v_i(t)\) by a chaotic system:

\[
\begin{align*}
\frac{y_j(t+1)}{v_j(t+1)} &= \begin{cases} 
2\text{sgn}(y_j(t))Th_{ij} - y_j(t) & \text{for } (v_j(t), y_j(t)) \in \Pi, \\
R & \text{otherwise},
\end{cases}
\end{align*}
\]

\[(5a)
\]

\[
\begin{align*}
\frac{y_j(t+1)}{v_j(t+1)} &= \begin{cases} 
0 & \text{for } (v_j(t), y_j(t)) \in \Pi, \\
\cos \theta & \text{otherwise},
\end{cases}
\end{align*}
\]

\[(5b)
\]

where \(y_j(t) = x_j(t) - \frac{1}{4}(pb_j + gb_j)\), \(Th_{ij} = \frac{1}{2}[pb_{ij} - gb_{ij}]\), \(\Pi_{ij} = \{(v_j(t), y_j(t)) | y_j(t) > Th_{ij}, \text{sgn}(v_j(t)y_j(t)) = -1\}\), \(\text{sgn}(a) = 1\) for \(a > 0\), \(-1\) otherwise, and \(R\) and \(\theta\) are system parameters. In this paper, \(R = 1.4\) and \(\theta = 50[\text{deg}]\) that guarantees Eq. (5) exhibits chaos [6] are selected. Because of the chaotic behavior, there is no diversity loss. The Algorithm of OPRC equals to Algorithm 1 except procedure 2. The second procedure is described in Algorithm 2.

2.4. Downhill simplex (DS)

DS tries to find the optima by \(d + 1\) searching points. The algorithm is described in Algorithm 3. DS updates a searching point \(x_i(t)\) with reflection, expansion and contraction procedure where \(w\) is index of the point obtaining the worst fitness at \(t\). With shrinking procedure, the searching points except best are updated. Therefore, DS always updates the searching points with improving \(f(x_i)\), then, the diversity loss could be taken place.

3. Tolerance memory update

In this section, a simple method is introduced to solve the outdated memory problem [5]. For PSO, APSO and OPRC, the update condition of \(pb_i\): \(f(x_i) > f(pb_i)\) (line 13 of Algorithm 1) and \(gb\): \(f(x_i) > gb\) (see also line 15) are replaced by \(f(x_i) > (f(pb_i) - \alpha)\) and \(f(x_i) > (gb - \alpha)\), respectively, where \(\alpha\) is tolerance parameter. For DS, the update condition of \(x_w\): \(f(x^w) > f(x_w)\) (line 11 of Algorithm 3), \(f(x^w) > f(x^*)\) (see also line 13), \(f(x^{**}) > f(x_b)\) (line 22), and \(f(x_w) > f(x^{**})\) (line 31) are replaced by \(f(x^w) > (f(x_w) - \alpha)\), \((f(x_w) + \alpha) > f(x^*)\), \((f(x^w) + \alpha) > f(x^{**})\), respectively. In this paper, \(\alpha = 5\).

4. Experiments

4.1. Moving peaks benchmark (MPB)

In this paper, performance of the algorithms is evaluated for moving peaks benchmark (MPB) [9, 10] as follows:

\[
f(x, n_e) = \max_p \left\{ H_p - W_p(n_e) \right\} \sum_{j=1}^{2} (x_j - x^*_p(n_e))^2
\]

(6)

where \(x \in \mathbb{R}^d\) is the evaluated position, \(p\) is index of the peak, \(H_p\) and \(W_p(n_e)\) is height and width of \(p\)-th peak, respectively, \(x^*_p(n_e) \in \mathbb{R}^d\) is maximum position of \(p\)-th peak. \(H_p\) is a static parameter given by uniformed distribution with [30, 70]. The \(W_p(n_e)\) and \(x^*_p(n_e)\) are updated by

\[
W_p(n_e) = \begin{cases} 
W_p(n_e - P) + \sigma & \text{for } n_e \ mod \ P = 0, \\
W_p(n_e - 1) & \text{otherwise},
\end{cases}
\]

\[
x^*_p(n_e) = \begin{cases} 
x^*_p(n_e - P) + V_p(n_e) & \text{for } n_e \ mod \ P = 0, \\
x^*_p(n_e - 1) & \text{otherwise},
\end{cases}
\]

\[
V_p(n_e) = \frac{1}{r + V_p(n_e - P)} \left( (1 - \lambda)r + AV_p(n_e - P) \right).
\]

where \(P\) is a period of change, \(\sigma\) is a random number of normally distribution with mean 0 and variation 1, \(\lambda\) is a inertia coefficient, \(l\) is distance of shift, and \(r\) is a random number.
Algorithm 3 Downhill Simplex

1: procedure Initialization
2: \( t = 0, n_e = 0 \)
3: \( x_i(t) \in \mathbb{R}^d, i = 1, \ldots, (d + 1) \)
4: for \( i = 1 \) to \( n \) do
5: \( f(x_i(t), n_e), n_e = n_e + 1 \)
6: \( w = \arg \min f(x_i), b = \arg \max f(x_i) \)
7: \( \bar{x} = \frac{1}{d} \sum_{i=1, i \neq w}^{d+1} x_i(0) \)
8: procedure Reflection
9: \( f^* = f(x^*, n_e), n_e = n_e + 1 \)
10: if \( f^* > f(x_b) \) then
11: Go to Expansion (procedure 3)
12: else
13: Go to Contraction (procedure 4)
14: end
15: procedure Expansion
16: \( x_w(t + 1) = x^*, f(x_w) = f^* \)
17: \( x_i(t + 1) = x_i(t), \text{where } i \neq w \)
18: Go to procedure 6
19: procedure Contraction
20: \( f^{**} = f(x^{**}, n_e), n_e = n_e + 1 \)
21: if \( f^{**} > f(x_b) \) then
22: \( x_w(t + 1) = x^{**}, f(x_w) = f^{**} \)
23: else
24: \( x_i(t + 1) = x^*, f(x_w) = f^* \)
25: \( x_i(t + 1) = x_i(t), \text{where } i \neq w \)
26: Go to procedure 6
27: procedure Shrinking
28: \( x^{**} = \bar{x} + \gamma(x^* - \bar{x}) \)
29: \( f^{**} = f(x^{**}, n_e), n_e = n_e + 1 \)
30: if \( f^{**} > f(x_b) \) then
31: Go to Shrinking (procedure 5)
32: else
33: \( x_w(t + 1) = x^{**}, f(x_w) = f^{**} \)
34: \( x_i(t + 1) = x_i(t), \text{where } i \neq w \)
35: Go to procedure 6
36: procedure Shrinking
37: \( x_i(t + 1) = \frac{1}{d}(x_i(t) + x_l(t)) \)
38: for \( i = 1 \) to \( d + 1 \) do
39: \( f(x_l) = f(x_i(t), n_e), n_e = n_e + 1 \)
40: Go to procedure 6
41: procedure Evaluation and Check Termination
42: \( w = \arg \min f(x_i), b = \arg \max f(x_i) \)
43: \( \bar{x} = \frac{1}{d} \sum_{i=1, i \neq w}^{d+1} x_i(t + 1) \)
44: if \( t + 1 = t_{\max} \) then
45: terminate
46: else
47: \( t = t + 1 \)
48: return to procedure 2

4.2. Behavior of tacking optima for 1-d MPB

In this section, tracking performance for single peak with 1-dimensional MPB is considered. The initial position \( x_{i1}(0) \) is given by \([x_{i1}^*(0) - 1, x_{i1}^*(0) + 1]\) with uniformed distribution, where \( g \) is the index that obtains the maximum \( H_p \). To compare under same condition, 1-dimensional MPB was generated once, and compared methods tried to track the generated optima for all trials. The tracking behaviors with 10 trials are shown in Fig. 1. As shown in Figs. 1a and 1b, PSO and OPRC could not track the solution because of the outlanded memory [5]. APSO and DS without the tolerance memory update also could not success the tracking. With the tolerance memory update, PSO could track the peak with several trials, however, it still lost the optima as shown in Fig. 1c. The same behavior is observed with DS. PSO and DS seems to be diversity loss [5], therefore, they can miss the solution even though their fitness memory are updated with tolerance. As shown in Figs. 1d and 1e, APSO and OPRC could track the optimal solution, and the tracking behavior of OPRC is qualitatively better than APSO because the diversity is low (see blue lines). This result suggests OPRC with the tolerance memory update solves the outlanded memory and the diversity loss issue without modification of the dynamical system.

4.3. Performance of tacking optima for MPBs

In this section, the performance of tracking optima is measured by mean error based on fitness value \( f_{err}(n_e) \) and mean error based on distance \( d_{err}(n_e) \), calculated by

\[ f_{err}(n_e) = \frac{1}{10} \sum_{m=1}^{10} \left( f(x_m^*(n_e), n_e) - f(x_{best}(m, n_e), n_e) \right), \quad (7) \]

\[ d_{err}(n_e) = \frac{1}{10} \sum_{m=1}^{10} \left( |x_m^*(n_e) - x_{best}(m, n_e)| \right), \quad (8) \]

where \( m \) is index of trials and \( x_{best}(m, n_e) \) is the searched best position at \( n_{e} \)-th evaluation of \( m \)-th trial. The experimental condition is common with Sec. 4.2: \( x_m^*(n_e) \) is common for all trials. The tolerance memory update is applied for all methods. As shown in table 1 and 2, OPRC obtained the lowest mean value and the lowest standard deviation of \( f_{err}(n_e) \) and \( d_{err}(n_e) \) for one and two dimensional MPB.

5. Conclusion

We have proposed tolerance memory update and applied it to OPRC. The method exhibited qualitatively better tracking behavior and obtained quantitatively low errors.
Table 1: Mean and standard deviation of $f_{err}(n_e)$ with tolerance memory update ($\alpha = 5$). The bold shows the best.

<table>
<thead>
<tr>
<th>dim</th>
<th>$f_{err}(n_e)$</th>
<th>PSO</th>
<th>APSO</th>
<th>OPRC</th>
<th>DS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>15.83</td>
<td>5.144</td>
<td><strong>1.594</strong></td>
<td>18.22</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>10.92</td>
<td>53.94</td>
<td><strong>1.359</strong></td>
<td>22.75</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>29.05</td>
<td>16.68</td>
<td><strong>3.922</strong></td>
<td>38.75</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>22.41</td>
<td>23.91</td>
<td><strong>1.666</strong></td>
<td>25.02</td>
</tr>
</tbody>
</table>

Table 2: Mean and standard deviation of $d_{err}(n_e)$ with tolerance memory update ($\alpha = 5$). The bold shows the best.

<table>
<thead>
<tr>
<th>dim</th>
<th>$d_{err}(n_e)$</th>
<th>PSO</th>
<th>APSO</th>
<th>OPRC</th>
<th>DS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean</td>
<td>5.017</td>
<td>1.134</td>
<td><strong>0.4136</strong></td>
<td>3.918</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>2.779</td>
<td>14.38</td>
<td><strong>0.3473</strong></td>
<td>3.838</td>
</tr>
<tr>
<td></td>
<td>mean</td>
<td>8.865</td>
<td>3.540</td>
<td><strong>0.6587</strong></td>
<td>20.27</td>
</tr>
<tr>
<td></td>
<td>std</td>
<td>6.823</td>
<td>15.43</td>
<td><strong>0.2809</strong></td>
<td>15.11</td>
</tr>
</tbody>
</table>

than PSO, APSO and DS with the tolerance memory update, respectively. Based on the results, it is confirmed that the proposed tolerance memory update can solve the outdated memory issue, and OPRC does not have the diversity loss. Therefore, OPRC can track the optimal solution using the tolerance memory update even though no detection methods and no modification of the dynamical system.

Considering the proposed method for the other benchmarks and real-world problems are future works.

References

Analysis of Resonator Modes in a Penrose Unilluminable Room

Takehiro Fukushima, Takao Sarata, Koichiro Sakaguchi, and Yasunori Tokuda

Department of Information and Communication Engineering, Okayama Prefectural University
111 Kuboki, Soja, Okayama 719-1197, Japan
Email: fuku@c.oka-pu.ac.jp

Abstract—Resonator modes in a Penrose unilluminable room are calculated using the finite-difference time-domain method. Spatially chaotic wavefunctions are obtained in the regions where the ray trajectories are confined. A chaotic wavefunction for which light spreads across the whole cavity was also derived. The existence of a widely spreading chaotic wavefunction shows that the properties of the unilluminable room derived from ray dynamical simulations do not always represent all of the resonator modes.

1. Introduction

The ray-wave correspondence in two-dimensional dielectric optical cavities is an interesting research topic from a viewpoint of fundamental physics [1–3]. In 1958, Penrose considered a two-dimensional cavity that always has dark regions which light rays cannot reach when a point light source is placed in the cavity [4,5]. This configuration is called a Penrose unilluminable room. Recently the ray dynamics in such a room have been investigated and have been found to exhibit three kinds of chaotic ray trajectories confined in different regions [6]. The light wave propagation has been modeled for a point light source using the finite-difference time-domain (FDTD) method [7]. A small amount of light leaks into the dark regions due to diffraction at the edges of the cavity. It would be interesting to investigate the modal properties of this optical cavity. In the present work, the resonator modes in a Penrose unilluminable room are calculated using the FDTD method.

2. Model

Figure 1 is a diagram of the Penrose unilluminable room studied here [7]. The left and right curved mirrors are half-ellipses whose semimajor and semiminor axes are 1.60 and 1.00 µm, respectively. Points F₁ and F₂ are the foci of the left half-ellipse, while points F₃ and F₄ are those of the right half-ellipse. The cavity has four arm regions labeled A, A', B, and B'. The top and bottom curved mirrors connecting points F₁ and F₃ and points F₂ and F₄ are half-ellipses whose semimajor and semiminor axes are 1.00 and 0.649 µm, respectively. The refractive index of the room is set to 3.3 to match that of GaAs laser diode cavity at a target wavelength of 860 nm. To ensure total internal reflection at the cavity edges, the region outside the cavity is taken to be a perfect electrical conductor.

Figure 1: Schematic of the Penrose unilluminable room.

The cavity has three kinds of chaotic ray trajectories [6]. One set of rays starts in region A or A', and is confined to regions A, P, and A' as illustrated in Fig. 2(a). Another set starts in region B or B', and is confined to regions B, Q, and B' as shown in Fig. 2(b). The third kind consists of trajectories starting in region M, and is confined to regions P, M, and Q as sketched in Fig. 2(c). The cavity also has three stable periodic orbits, namely the axial orbit in Fig.

Figure 2: Ray trajectories confined in the cavity: (a) chaotic ray trajectories in regions A, P, and A'; (b) chaotic ray trajectories in regions B, Q, and B'; (c) chaotic ray trajectories in regions P, M, and Q; (d) a stable axial periodic orbit; (e) a stable diamond-shaped orbit; and (f) two stable V-shaped orbits.
2(d), the diamond-shaped orbit in Fig. 2(e), and two V-shaped orbits in Fig. 2(f).

The commercial software package FullWAVE [8] is used for the FDTD calculations. To obtain resonator modes associated with the chaotic ray trajectories, an impulsive light source having a Gaussian beam profile and a point monitor of the temporal waveform are positioned inside the cavity. The waveform of the magnetic field is measured and its spectrum is calculated from a Fourier transform. The mode patterns of the magnetic field distribution inside the cavity at the resonance wavelengths are calculated. Two combinations are used for the positions of the light source and waveform monitor. Combination 1 consists of the light source in region A (at $x = -0.65 \mu m$ and $z = -1.4245 \mu m$) with the waveform monitor in region P (at $x = -1.6 \mu m$ and $z = 0$) to excite resonator modes associated with the chaotic ray trajectories of Fig. 2(a). Combination 2 has the light source in region M (at $x = 0.1 \mu m$ and $z = 0.2 \mu m$) with the waveform monitor in region P (at $x = -1.6 \mu m$ and $z = 0$) to excite resonator modes associated with the chaotic ray trajectories of Fig. 2(c). The impulsive light is emitted in the negative x direction in both cases. The light propagating in the cavity is $p$-polarized with electromagnetic components $E_x$, $H_y$, and $E_z$.

3. Results

Figure 3 plots the $H_y^2$ spectrum for combination 1 in the wavelength region from 800 to 900 nm. Several resonance peaks can be seen. Figure 4 shows the magnetic field distribution for four resonator modes corresponding to the peaks found in Fig. 3. Spatially chaotic wavefunctions confined to regions A, P, and A' are shown in Fig. 4(a), (b), and (c). A small amount of light leaks into region M in Fig. 4(b). A spatially chaotic wavefunction spreading across the whole cavity is shown in Fig. 4(d). In the ray dynamical simulations, optical rays starting in region A are completely confined to regions A, P, and A' and can never reach regions M, Q, B, and B', as seen in Fig. 2(a). In the modal analysis, on the other hand, the light is evidently not strictly confined to regions A, P, and A'.

Figure 5 graphs the $H_y^2$ spectrum for combination 2 in the wavelength region from 800 to 900 nm. The four peaks indicated by the blue arrows correspond to resonator modes quantized along the stable axial periodic orbit, as shown in Fig. 6. We also found higher order axial resonator modes, as shown in Fig. 7(a), and complex spatially chaotic wavefunctions that spread into regions P, M, and Q, as shown in Fig. 7(b). The chaotic wavefunctions are associated with the chaotic ray trajectories of Fig. 2(c). However, a small amount of light leaks into regions A, A', B, and B'.

The Penrose unilluminable room has three kinds of chaotic ray trajectories confined to different regions. These trajectories are independent of each other. In the modal analysis based on the FDTD method, some spatially chaotic wavefunctions arise that are associated with the chaotic ray trajectories. Moreover, a chaotic wavefunction is found that spreads across the whole cavity. This widely spreading

![Figure 3: Spectrum of $H_y^2$ for combination 1.](image)

![Figure 5: Spectrum of $H_y^2$ for combination 2.](image)

![Figure 4: Magnetic field patterns for resonator modes at the indicated wavelengths.](image)
wave implies that coupling exists among the three kinds of chaotic ray trajectories. In previous work, it was shown that a small amount of light leaks into the dark regions due to diffraction at the cavity edges [7]. The coupling among the three chaotic ray trajectories may also be caused by diffraction.

4. Conclusions

Resonator modes have been calculated for a Penrose unilluminable room using the FDTD method. Spatially chaotic wavefunctions associated with three kinds of chaotic ray trajectories have been found. Moreover, a chaotic wavefunction has been discovered that spreads across the whole cavity, implying coupling among the three chaotic ray trajectories. This coupling may arise from diffraction at the cavity edges. The Penrose unilluminable room is expected to have a wide variety of resonator modes in addition to those obtained here. Thus, the correspondence between rays and waves in this type of optical cavity may reveal new aspects of fundamental physics.

Acknowledgment

T. F. first learned about Penrose unilluminable room from Prof. A. E. Siegman in 2007. The authors thank Mr. Yoshiki Mihara for assistance with the numerical calculations.

References

Wave-chaos-induced single-frequency lasing in microcavities

Satoshi Sunada¹, Susumu Shinohara², Takehiro Fukushima³, and Takahisa Harayama⁴

¹Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University
Kakuma-machi Kanazawa, Ishikawa 920-1192, Japan
²NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai Seika-cho Soraku-gun, Kyoto 619-0237, Japan
³Department of Information and Communication Engineering, Okayama Prefectural University
111 Kuboki Soja, Okayama 719-1197, Japan
⁴Department of Applied Physics, School of Advanced Science and Engineering, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
Email: sunada@se.kanazawa-u.ac.jp

Abstract—We experimentally and numerically demonstrate that stable single-frequency (i.e., single-mode) lasing can be achieved in microcavities that have wave-chaotic modes. The single-frequency lasing is explained as a result of strong competitive interactions among wave-chaotic modes.

1. Introduction

Various active devices ranging from musical instruments to lasers generate oscillating states with well-defined frequencies from the interplay between resonator geometry and an active nonlinear element [1, 2]. Understanding and controlling the formation of such self-organized oscillating states is important in device physics and related applications. As a specific example, two-dimensional (2D) microcavity lasers have attracted considerable attention over the past decades [3, 4, 5]. Depending on the cavity shapes, they can exhibit a variety of lasing states through the interaction between the light field and active gain material [2]. The studies of 2D microcavity lasers have led to a wide range of applications, including low-threshold microlasers with unidirectional emission [6], low-coherence microlasers [7], and fast random signal generation [8]. Moreover, 2D microcavity lasers have served as a platform for experimentally addressing fundamental issues such as quantum/wave chaos in open systems [5] and non-Hermitian physics [9, 10].

The lasing emission patterns and spectra are fundamental and important characteristics of 2D microcavity lasers. Previous experimental and theoretical studies have shown that the lasing emission patterns can be well characterized by resonant modes, which are determined only by the cavity shape and refractive index [3, 4]. Meanwhile, the spectral characteristics are still poorly understood because nonlinear interaction among resonant modes plays an crucial role in determining lasing frequencies.

In this presentation, we report our recent finding on a relationship between shapes of 2D microcavities and the lasing frequencies [11]. We show that stable single-frequency (i.e., single-mode) lasing can be achieved in fully chaotic cavities, where all of the internal ray orbits are chaotic in a ray optics picture and the spatial patterns of the modes are wave-chaotic, whereas multi-frequency (i.e., multimode) lasing is exhibited in non-chaotic cavities. The achievement of single-mode lasing in fully chaotic cavities is explained as a result of strong mode competition among wave chaotic modes.

2. Fully chaotic and non-chaotic microcavities

A specific shape of the fully chaotic cavity that we study here is a stadium [12], which is widely used for classical and quantum chaos studies. As shown in Fig. 1(a), the stadium cavity consists of two straight lines of length l and two half circles of radius R. We define the aspect ratio parameter \( p = W/L \), where \( W = 2R \) is the length of the minor axis and \( L = 2R + l \) is the length of the major axis. For the non-chaotic cavity, we focus on the elliptic cavity defined in Fig. 1(b). It is known that a closed elliptic cavity is an integrable system [13], thus exhibiting no chaotic behavior. In the same manner as for the stadium cavity, we define the aspect ratio parameter for the elliptic cavity as \( p = B/A \), where \( A \) and \( B \) are the lengths of the major and minor axes, respectively.

We fabricated semiconductor microcavities with the stadium and elliptic cavities by applying a reactive-ion-etching technique to a graded index separate-confinement-heterostructure (GRIN-SCH) single-quantum-well GaAs/Al,Ga1−xAs structure grown by MOCVD (See Ref. [14] for details on the layer structures and fabrication process). The fabricated lasers are shown in Figs. 1(c) and 1(d). In our experiments, the lasers were soldered onto aluminum nitride submounts at 20 ± 0.1 °C and electrically driven with cw current injection. The optical outputs were collected with anti-reflection-coated lenses and coupled to a multimode optical fiber via a 30-dB optical isolator.
3. Experimental results

Figure 2 shows typical lasing wavelength spectra for the stadium cavity laser and elliptic cavity lasers. Because the lasing wavelength $\lambda$ is related to a frequency $f$ by $c = f\lambda$, where $c$ is the light velocity, we discuss the spectral characteristics in the wavelength regime. As shown in Fig. 2, there is a remarkable difference in the number of peaks between the two lasers. The spectrum of the stadium cavity laser exhibits only a single sharp peak, suggesting single frequency (single-mode) lasing, despite the fact that the number of modes within the gain band is more than a few thousand because of the large cavity area. Meanwhile, the spectrum of the elliptic cavity laser always exhibits multiple peaks, i.e., multimode lasing.

We systematically investigated the spectral characteristics of the two lasers with various cavity areas $S$ and aspect ratios $p$. Figure 3 shows the number of peaks whose intensities were larger than $-20$ dB of the maximum peak intensity in each spectrum as a function of the current $I$ normalized by the threshold current $I_{th}$ for each laser. The spectra of the elliptic cavity lasers always exhibit multiple peaks, i.e., multimode lasing, and the number of peaks increases as $I$ increases. On the other hand, the spectra of the stadium cavity lasers always exhibit a single peak, i.e., single-mode lasing, regardless of the cavity area and aspect ratios, and the single-mode lasing is maintained even for high injection current values $I$. However, as a slight exception, we observed two peaks for the stadium cavity laser for $I/I_{th} = 2.1$ and 2.7 in Fig. 3. We attribute this to mode hopping caused by a thermal effect of the current injection, such as a gain shift and a change in the refractive index.

4. Discussion and Analysis

The above results indicate that strong suppression of multimode lasing is a common feature of stadium cavity
lasers. As discussed in Refs. [15, 16], a competitive interaction occurs among modes that are overlapped not only spectrally but also spatially. In a fully chaotic cavity, modes typically have complex spatial patterns that spread throughout the entire cavity due to the ray dynamical property [for a typical example, see Fig. 4(a)], and therefore result in large spatial overlaps with other modes. Actually, previous numerical simulations of stadium cavity lasers demonstrated a strong selection of lasing modes owing to a competitive interaction [17, 18, 19]. On the other hand, non-chaotic cavities typically support spatially localized modes [e.g., Fig. 4(b)]. Interestingly, different modes are localized in different areas. Thus, when the spatial overlap among modes is small, the competition among them can be avoided. Indeed, the simultaneous lasing of multiple modes for a non-chaotic cavity laser with a circular shape was numerically demonstrated in Ref. [20].

To quantify the spatial overlaps between two modes in a cavity, we introduce the following cross-correlation for the amplitude distributions of resonance modes:

\[ C = \frac{\int |\phi(r)| |\psi(r)| w(r) dr}{\sqrt{\left( \int |\phi(r)|^2 w(r) dr \right) \left( \int |\psi(r)|^2 w(r) dr \right)}}, \tag{1} \]

where \( \phi(r) \) and \( \psi(r) \) are the modal wave functions and \( w(r) \) represents a pumping region. For uniform pumping, \( w(r) = 1 \) inside the cavity, whereas \( w(r) = 0 \) outside. The correlation is essentially similar to a spatial contribution to the cross-gain saturation (i.e., intensity cross-correlation) between two modes [15, 16, 21]. Using the boundary element method [22], we calculated the resonances of the stadium and elliptic cavities with \( p = 0.5 \), imposing a refractive index of 3.3 inside the cavities and transverse electric (TE) polarization. Because of computational power limitations, we set a size parameter \( 2\pi R/\lambda \approx 100 \), where \( R \) and \( \lambda \) are the characteristic radius and wavelength, respectively. This size parameter value is smaller than that of a real laser cavity used in the experiments but is sufficiently large to discuss the properties of the wave functions in the short-wavelength regimes [23]. We obtained approximately 100 low-loss modes with a quality factor \( Q \geq 3000 \) for the stadium cavity, whereas \( Q \geq 3 \times 10^3 \) for the elliptic cavity. Typical examples of the wave functions for stadium and elliptic cavities are shown in Fig. 4.

Figure 5 shows the histogram of the spatial overlap \( C \) between two low-loss modes in stadium and elliptic cavities. The \( C \)-values for the stadium cavity are distributed around 0.77. In contrast, the \( C \)-values for the elliptic cavity are widely distributed with a mean value of 0.45. The relatively low \( C \)-values come from the small spatial overlaps between the localized modes. In particular, the overlaps are small between two modes with different radial mode numbers \( n_r \), which characterize the number of field maxima in the radial direction. For instance, the \( C \)-value between the modes with \( n_r = 5 \) and \( n_r = 1 \) was only 0.14. As seen in Fig. 2(b), the lasing peaks in the spectra of the elliptic cavity laser are not always equally spaced. This result means that modes with different \( n_r \)-values were involved in the lasing, which supports our interpretation of the relation between the spatial overlaps and the spectral characteristics.

5. Conclusion

We experimentally investigated the difference in the spectral characteristics between fully chaotic cavity lasers with a stadium shape and non-chaotic cavity lasers with an elliptic shape. In the stadium cavity lasers, only a single mode was excited at high pumping regimes regardless of the size and aspect ratio, whereas many modes were excited in the elliptic cavity lasers. The strong suppression of multimode lasing observed in the stadium cavity lasers can
be explained by the large spatial overlaps among the low-loss modes. Because a common feature of modes in a fully chaotic cavity is the spatial pattern that spreads throughout the entire cavity, we expect that the modal suppression leading to single-mode lasing is a universal feature of fully chaotic cavity lasers.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers 26790056 and 16K04974.

References

Asymmetric emission caused by chaos-assisted tunneling and synchronization in two-dimensional microcavity lasers

Yuta Kawashima¹, Susumu Shinohara², Satoshi Sunada³, Takehiro Fukushima⁴ and Takahisa Harayama⁴

¹ Department of Applied Physics, School of Advanced Science and Engineering, Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
² NTT Communication Science Laboratories, NTT Corporation, 2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
³ Faculty of Mechanical Engineering, Institute of Science and Engineering, Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan
⁴ Department of Information and Communication Engineering, Okayama Prefectural University, 111 Kuboki, Soja, Okayama 719-1197, Japan
Email: louzoulbouce@asagi.waseda.jp

It has been experimentally reported that low-threshold lasing and directional emission can be simultaneously achieved in a two-dimensional microlaser with the quadrupolar cavity [1]. The quadrupolar cavity is defined in the polar coordinates \((r, \theta)\) as
\[
r(\theta) = r_0(1 + \epsilon \cos 2\theta),
\]
where \(r_0\) is the size parameter and \(\epsilon\) is the deformation parameter fixed as \(\epsilon = 0.09\). In this cavity, a resonant mode that localizes along a pair of stable triangular orbits is confined by total internal reflection. This strong light confinement leads to a low lasing threshold, while weak light emission occurs by chaos-assisted tunneling [2]. The tunneling induces light intensity leakage from the stable triangular orbits to chaotic orbits that eventually escape from the cavity violating the critical angle condition for total internal reflection, where the chaotic dynamics governed by unstable manifolds results in directional emission [3].

The emission patterns of two-dimensional microcavity lasers can be theoretically studied by analyzing the resonant modes of the two-dimensional Helmholtz equation [4]
\[
\nabla^2 + n^2(x, y) k^2 \psi(x, y) = 0,
\]
where \(n(x, y)\) is the refractive index and \(k\) is the vacuum wave number. The real part of \(k\) represents the resonant wave number, while the imaginary part represents the decay rate of the resonant mode. We consider transverse-magnetic polarization, namely, \(\psi(x, y)\) represents the \(z\)-component of the electric field \(E_z\). The refractive index inside the cavity is \(n_{in} = 3.3\), while it is \(n_{out} = 1\) outside the cavity. Equation (2) can be numerically solved by, for example, the extended boundary element method [5].

In the experiment in Ref. [1], asymmetric directional emission patterns were observed despite the symmetry of the quadrupolar cavity. The wave functions of the resonant modes are divided into the four symmetry classes
\[
\psi_{ab}(-x, y) = a \psi_{ab}(x, y) \quad a, b \in \{-+, +\}
\]
with the parities \(a, b \in \{-+, +\}\). Therefore the emission patterns of individual resonant modes have the same symmetry as the cavity shape (i.e., symmetric with respect to both \(x\) and \(y\) axis). However, when we consider nonlinear modal interaction through a lasing medium, the locking, or synchronization of two different parity modes can occur, and it yields an asymmetric emission pattern [6].

In this presentation, we show that the experimentally observed asymmetric emission pattern can be explained by the locking of a nearly degenerate pair, \(\psi_{+-}\) and \(\psi_{-+}\) (i.e., the modes with slightly different wave numbers). First, we demonstrate that although \(\psi_{+-}\) and \(\psi_{-+}\) are localized along the pair of stable triangular orbits, their superposition turns out to be localized only along one of the pair. Secondly, by using the Maxwell-Bloch model that takes account the nonlinear effect of a lasing medium [7], we actually simulate the locking of the nearly degenerate pair as well as the appearance of the asymmetric emission patterns.

References

Chaos synchronization and nonlinear dynamics in a photonic integrated circuit with two semiconductor lasers

Shoma Ohara¹, Andreas Karsaklian Dal Bosco¹, Kazusa Ugajin¹, Atsushi Uchida¹, Takahisa Harayama²,³ and Kazuyuki Yoshimura²,⁴

¹ Department of Information and Computer Sciences, Saitama University
255 Shimo-Okubo Sakura-ku, Saitama City, Saitama 338-8570, Japan
² NTT Communication Science Laboratories, NTT Corporation
3-1 Morinosato, Wakamiya, Atsugi-Shi, Kanagawa 243-0198, Japan
³ Department of Applied Physics, Waseda University
3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
⁴ Department of Information and Electronics, Graduate school of Engineering, Tottori University
4-101 Koyama-Minami, Tottori 680-8552, Japan

Emails: s16dm002@mail.saitama-u.ac.jp, auchida@mail.saitama-u.ac.jp

Abstract—We investigate the frequency dependence of chaos synchronization in a photonic integrated circuit with mutually-coupled semiconductor lasers. We calculate the cross-correlation between the two temporal waveforms of the two laser outputs with a low-pass filter. In-phase synchronization is observed between chaotic temporal waveforms of the two lasers. On the contrary, anti-phase synchronization is observed when the lasers exhibit low-frequency fluctuations (LFF).

1. Introduction

Coupled nonlinear systems show a large variety of dynamics. Recently, chaos synchronization in semiconductor lasers with delayed optical feedback has been studied for applications in secure key distribution [1, 2]. When two semiconductor lasers are mutually coupled, the output of these lasers show chaotic temporal oscillations whose dominant frequency corresponds to the inverse of twice the coupling time delay [3]. The chaotic temporal waveforms can be synchronized to each other with the time lag of the coupling delay time. In addition, anti-synchronization of low-frequency fluctuations (LFF) has been observed [4]. Episodic synchronization also has been reported when optical frequency detuning is changed [5].

The dynamics and synchronization in mutually-coupled semiconductor lasers have been investigated intensively for long coupling lengths (> 100 mm). On the contrary, photonic integrated circuits (PICs) have been proposed recently as monolithically integrated optical systems suitable for physical random number generation [6, 7]. However, the study of chaos synchronization for short coupling lengths in mutually coupled lasers has few reports. Synchronization has been performed in a particular case of two lasers exhibiting periodic oscillations in a photonic integrated circuit [8]. Nevertheless, nonlinear dynamics and chaos synchronization in a PIC with two mutually-coupled semiconductor lasers with short coupling length (~ 10 mm) have not been reported yet. It is important to investigate chaos synchronization and nonlinear dynamics in a photonic integrated circuit with mutually-coupled semiconductor lasers.

In this study, we investigate chaos synchronization in a PIC with two mutually-coupled semiconductor lasers. We focus on the dependence of the synchronization quality on different frequency components by using a low-pass filter in the LFF regime.

2. Experimental setup

We present the configuration of our photonic integrated circuit with mutually-coupled semiconductor lasers in Fig. 1. In this PIC, two semiconductor lasers, two photodetectors, a semiconductor optical amplifier (SOA), and an external mirror are monolithically integrated. The lasers are mutually coupled via the external mirror. In addition, each laser is subjected to its own optical feedback, and the corresponding external cavity lengths are 11.0 mm for laser 1 and 10.3 mm for laser 2, respectively. The parameters of the PIC are the injection currents for the laser 1, and laser 2, whose lasing threshold currents are 12.0 mA. In addition, we can change the feedback strength of laser 2 and the coupling strength between the two lasers through the SOA injection current.

Figure 1: Schematics of photonic integrated circuit.
3. Frequency dependence of synchronization in LFF regime

The LFF dynamics consists of high-frequency chaotic oscillations and low-frequency intensity dropouts [9, 10]. We apply a low-pass filter to the laser output signals to separate these two dynamics. We calculate the cross correlation between the temporal waveforms of the laser 1 and 2 after filtering of the two laser outputs to evaluate the synchronization quality for different cut-off frequencies (1 GHz and 16 GHz) of the low-pass filter.

Figure 2 shows the temporal waveforms and the correlation plots of the output of the two lasers when the cut-off frequencies of the low-pass filter are set to 16 GHz (top) and 1 GHz (bottom), respectively. When the signals are filtered at 16 GHz, the lasers show in-phase synchronization. However, the lasers show anti-phase synchronization when the 1 GHz filter is applied. We found that the cross-correlation between the two lasers indicates a negative value for the filtered signals, while this value is positive for the unfiltered signals.

Figure 3 shows the dynamics and synchronization state between both lasers when the SOA injection current ($I_{SOA}$) is changed. We start from a low value of $I_{SOA}$ of 6.00 mA (Fig. 3(a)(d)), for which both laser 1 and laser 2 exhibit chaos without low-frequency predominance. When $I_{SOA}$ is increased to 25.00 mA (Fig. 3(b)(e)), Laser 1 still exhibits chaos, while Laser 2 enters a LFF regime, as seen from the increase of low-frequency components in the RF spectrum. When $I_{SOA}$ is further increased to 39.00 mA, both lasers exhibit LFF dynamics. Thus, changing the SOA injection current induces a change in their dynamics.

Figure 4 shows the evolution of the peak of the cross-correlation value when the SOA injection current is changed, filtered at 1 and 16 GHz. This figure corresponds to the results of Fig. 3. We also calculate the maximum of the absolute value of the cross-correlation value for each signal because the delay time indicating the peak value changes when the cut-off frequency is changed. We discuss the dependence of the synchronization of LFF dynamics between the two lasers on the cut-off frequency of the low-pass filter. The change in the coupling strength between the two lasers results in the change in their temporal dynamics. When chaos content is dominant in both lasers (0 mA $\leq I_{SOA} \leq 10$ mA), in-phase oscillations are observed for both cases of 16 GHz and 1 GHz filters. When low-frequency content is dominant in both lasers (31 mA $< I_{SOA} \leq 50$ mA), anti-phase synchronization are obtained. When low-frequency content is dominant in one of the two lasers (21 mA $\leq I_{SOA} \leq 31$ mA), anti-phase synchronization is observed when filtered at 16 GHz while anti-phase synchronization is obtained when filtered at 1 GHz. The synchronization state is dependent on the dynamics of each laser. Therefore, we understand that the low-frequency components of
LFF influence anti-phase synchronization.

Figure 4: Cross-correlation value when the SOA injection current is changed.

4. Numerical simulations

We investigate numerical simulations with the rate equations known as the Lang-Kobayashi equations [11] in order to reproduce the experimental results as well as to give theoretical explanation. The Lang-Kobayashi equations are written as follows:

\[
\frac{dE_{1,2}(t)}{dt} = \frac{1 + i\alpha}{2} \left( G_N(N_{1,2}(t) - N_0) \right) - \frac{1}{\tau_p} E_{1,2}(t) + \kappa_{1,2} E_{1,2}(t - \tau_{1,2}) \exp(-i\omega_{1,2} \tau_{1,2}) + \kappa_{in} E_{1,2}(t - \tau_{in}) \exp[i(\Delta \omega \tau_{in} - \omega \tau_{in})]\]

\[
\frac{dN_{1,2}(t)}{dt} = J_{1,2} - \frac{N_{1,2}(t)}{\tau_s} - \frac{G_N(N_{1,2}(t) - N_0)}{1 + \epsilon |E_{1,2}(t)|^2} |E_{1,2}(t)|^2 \]

Where \( E \) and \( N \) are the complex electric field and the carrier density, respectively. \( \tau_{1,2} \) and \( \kappa \) represent the feedback delay time and strength. \( \tau_{inj} \) and \( \kappa_{inj} \) represent the coupling delay time and strength. \( \alpha \) is the linewidth enhancement factor. \( J \) is the laser injection current, \( G_N \) is the gain coefficient, \( N_0 \) is the carrier density at transparency, \( \tau_p \) and \( \tau_s \) are the photon and carrier lifetimes. \( \epsilon \) is the gain saturation coefficient. \( \Delta \omega \) is the detuning of the optical angular frequencies between the two lasers. The parameter values are set as follows: \( J_1 = 1.02 \ J_{th}, J_2 = 1.10 \ J_{th}, k_1 = 0.349, k_2 = 0.099, k_{inj} = 0.067, \tau_1 = 0.29 \ ns, \tau_2 = 0.27 \ ns, \) and \( \tau_{inj} = 0.28 \ ns. \)

We investigate the synchronization state on different frequency components by using the low-pass filter. Figure 5 shows the temporal waveforms, correlation plots, and RF spectrum for the outputs of both lasers. The cut-off frequencies are set to 16 GHz and 1 GHz in Fig. 5(a)(b) and Fig. 5(c)(d), respectively. In-phase synchronization is observed at high-frequency components, however, anti-phase synchronization is observed at low-frequency components. We focus on the RF spectrum in Fig. 5(e). The high-frequency components are dominant for laser 1, while the low frequency components are dominant for laser 2, similar to Fig. 3(b)(e). Therefore, the numerical result agrees well with the experimental result of Fig. 2 and 3.

5. Conclusions

We investigated chaos synchronization in a photonic integrated circuit with two mutually-coupled semiconductor lasers. We applied a low pass filter with the cut-off frequency of 16 GHz or 1 GHz to the laser output signals to separate these two dynamics. We observed in-phase syn-
chronization at high-frequency components and anti-phase synchronization at low-frequency components. This result reveals the frequency dependence of chaos synchronization in the photonic integrated circuit with two mutually-coupled lasers. This phenomenon can be observed for intermediate level of the coupling strength and asymmetric optical-feedback for the two semiconductor lasers. The numerical results agree well with the experimental results.

Acknowledgments

We acknowledge support from Grants-in-Aid for Scientific Research from Japan Society for the Promotion of Science (JSPS KAKENHI Grant Number JP24686010), and Management Expenses Grants from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

References


Common-signal-induced synchronization in photonic integrated circuits driven by constant-amplitude random-phase light

Takuma Sasaki\textsuperscript{1}, Izumi Kakesu\textsuperscript{1}, Atsushi Uchida\textsuperscript{1}, Satoshi Sunada\textsuperscript{2,3}, Kazuyuki Yoshimura\textsuperscript{1,4} and Kenichi Arai\textsuperscript{2}

\textsuperscript{1}Department of Information and Computer Sciences, Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama City, Saitama, 338-8570, Japan
\textsuperscript{2}NTT Communication Science Laboratories, NTT Corporation 3-1 Morinosato, Wakamiya, Atugi-Shi, Kanagawa 243-0198, Japan
\textsuperscript{3}Faculty of Mechanical Engineering, Kanazawa University, Kakuma-machi, Kanazawa, Ishikawa 920-1192, Japan
\textsuperscript{4}Department of Information and Electronics, Graduate school of Engineering, Tottori University 4-101 Koyama-Minami, Tottori 680-8552, Japan

Emails: s15mm312@mail.saitama-u.ac.jp, auchida@mail.saitama-u.ac.jp

Abstract—We experimentally and numerically investigate common-signal-induced synchronization in photonic integrated circuits (PICs) driven by constant-amplitude random-phase light. We measure the cross-correlation value between the outputs of the two PICs when the feedback phase is changed. The temporal waveforms of PICs show high cross-correlation when the feedback phases are matched, whereas low correlation is obtained when they are mismatched. The RF spectra of two PICs are similar to each other when the feedback phase is matched.

1. Introduction

Information-theoretic security [1] has been studied as a new information-security paradigm to replace the computational security. A private key distribution method has been proposed using correlated random bits as a key distribution based on information theory security [2–5]. This method generates a secret key from a random number sequence with a correlation that can be shared by two users. Common-signal-induced synchronization is a key technique for this method and has been demonstrated using semiconductor lasers [6–10]. It utilizes the phenomenon that the output waveforms of two lasers (called Response lasers) are synchronized by injecting a common drive laser output. Synchronization is achieved only when the phase parameters between the Response lasers are matched. However, synchronization is unstable in an optical-fiber-based system with long external cavity length (e.g., several meters) [8] because the optical phase fluctuates under the influence of air turbulence and temperature fluctuation.

In this study, we experimentally and numerically investigate common-signal-induced synchronization in photonic integrated circuits driven by constant-amplitude random-phase (CARP) light.

2. Experimental setup

The structure of the photonic integrated circuit used in this study is shown in Fig. 1. The output of the distributed-feedback (DFB) laser is reflected by an external mirror. The phase and intensity of the optical feedback is adjusted by a phase modulator (PM) and an optical amplifier (SOA), respectively. The optical output of the DFB laser can be detected through an optical fiber, and CARP light can be injected as well. The external cavity length is 8.6 mm, which corresponds to the external cavity frequency of 4.9GHz. When the external cavity frequency is higher than the relaxation oscillation frequency, the laser is defined as short cavity regime (SCR). We set the injection current as $J = 1.2J_{th}$ for synchronization. In this case, the relaxation oscillation frequency is 1.8 GHz, which satisfies the condition of the short cavity regime with the PICs.

Figure 1: Schematics of the photonic integrated circuit. SOA, semiconductor optical amplifier; PM, phase modulator;

The experimental setup is shown in Fig. 2. We use a semiconductor laser (Drive) and two PICs (PIC 1, and PIC 2) for common-signal-induced synchronization. We generate an optical noise signal from the output of a superluminescent diode (SLD). The optical phase of the drive signal is modulated randomly by the output of the SLD to generate CARP light. The CARP light from the Drive laser is divided by a fiber coupler (FC). Each CARP light is injected into the PIC unidirectionally through an optical isolator (ISO). The injection strength of the CARP light is adjusted by an optical attenuator (ATT). The optical feedback of the PICs is controlled by the SOA current (called...
closed-loop configuration). The phase of the optical feedback in each PIC is modulated by a waveform generator. The output waveforms of the CARP light and the two PICs are observed by photodetectors and amplified by electric amplifiers. The converted electrical signals are detected by using a digital oscilloscope and a radio-frequency (RF) spectrum analyzer.

3. Experimental results

We experimentally investigate common-signal-induced synchronization when the photonic integrated circuits are used as the Response lasers. We introduce a measure of cross-correlation to evaluate the quality of synchronization. The cross-correlation value is calculated as follows:

\[ C = \frac{\langle (I_1 - \bar{I}_1)(I_2 - \bar{I}_2) \rangle}{\sigma_1 \cdot \sigma_2} \]  
(1)

where \( I_1, I_2 \) are the temporal waveforms of the output intensities of the PIC 1 and 2, respectively, \( \bar{I}_1, \bar{I}_2 \) are their mean values, \( \sigma_1, \sigma_2 \) are their standard deviations of \( I_1, I_2 \), and \( \langle \rangle \) is time averaging. \( C = 1 \) indicates identical synchronization, whereas \( C = 0 \) indicates no synchronization.

Figure 3 shows the temporal waveforms of the PIC 1 and 2 and their correlation plots. In Figs. 3(a) and (b), when the voltages of the PM are set to 0.0 V for both of the PICs, the temporal waveforms of the two PICs are strongly correlated. We obtain a high cross-correlation value of 0.932 as shown in Fig. 3(b). On the other hand, when the feedback phase is mismatched, the temporal waveforms of the two PICs are strongly correlated in Figs. 3(c) and (d). The cross-correlation shows a low value of 0.135 as shown in Fig. 3(d). We thus experimentally achieve common-signal-induced synchronization of the two PICs with phase modulation.

Figure 4 shows the RF spectra of the PIC 1 and 2. \( V_{PM1,2} \) denote the voltages applied to the PM of the PIC 1 and 2, respectively. The feedback phases of the PICs are matched in Fig. 4(a), while they are mismatched in Fig. 4(b). The two RF spectra are in good agreement in Fig. 4(a) with the same parameter values of the feedback phases. We change the voltage of the PM of PIC 1 to 1.176 V in Fig. 4(b) and the two RF spectra are not matched. The change in the RF spectra is observed by varying the feedback phase in the short-cavity regime due to the large external cavity frequency of PICs (4.9 GHz).

4. Numerical model

We conduct numerical simulations to reproduce our experimental results. We use the Lang-Kobayashi equations [11] to describe a model consisting of the PICs with CARP light injection as follows [7].

\[ \frac{dE_r(t)}{dt} = \frac{1 + ia}{2} \left[ G_N(N_r(t) - N_0) - \frac{1}{\tau_p} E_r(t) \right] + \kappa_r E_r(t - \tau_r) \exp[i(\omega_r \tau_r)] + \sigma E_d(t - \tau_{inj}) \exp[i(\omega_d t - \omega_r \tau_{inj})] \]  
(2)

The cross-correlation value is calculated as follows:

\[ C = \frac{\langle (I_1 - \bar{I}_1)(I_2 - \bar{I}_2) \rangle}{\sigma_1 \cdot \sigma_2} \]  
(1)

where \( I_1, I_2 \) are the temporal waveforms of the output intensities of the PIC 1 and 2, respectively, \( \bar{I}_1, \bar{I}_2 \) are their mean values, \( \sigma_1, \sigma_2 \) are their standard deviations of \( I_1, I_2 \), and \( \langle \rangle \) is time averaging. \( C = 1 \) indicates identical synchronization, whereas \( C = 0 \) indicates no synchronization.

Figure 3 shows the temporal waveforms of the PIC 1 and 2 and their correlation plots. In Figs. 3(a) and (b), when the voltages of the PM are set to 0.0 V for both of the PICs, the temporal waveforms of the two PICs are strongly correlated. We obtain a high cross-correlation value of 0.932 as shown in Fig. 3(b). On the other hand, when the feedback phase is mismatched, the temporal waveforms of the two PICs are strongly correlated in Figs. 3(c) and (d). The cross-correlation shows a low value of 0.135 as shown in Fig. 3(d). We thus experimentally achieve common-signal-induced synchronization of the two PICs with phase modulation.

Figure 4 shows the RF spectra of the PIC 1 and 2. \( V_{PM1,2} \) denote the voltages applied to the PM of the PIC 1 and 2, respectively. The feedback phases of the PICs are matched in Fig. 4(a), while they are mismatched in Fig. 4(b). The two RF spectra are in good agreement in Fig. 4(a) with the same parameter values of the feedback phases. We change the voltage of the PM of PIC 1 to 1.176 V in Fig. 4(b) and the two RF spectra are not matched. The change in the RF spectra is observed by varying the feedback phase in the short-cavity regime due to the large external cavity frequency of PICs (4.9 GHz).
\[
dN_r(t) = J_r - \frac{N_r(t)}{\tau_r} - \frac{G_N(N_r(t) - N_0)}{1 + |E_r(t)|^2} |E_r(t)|^2 \quad (3)
\]

\[
E_d(t) = \sqrt{I(\tau)} \exp[i\varphi(t)] \quad (4)
\]

Where, \(E_r(t)\) and \(N_r(t)\) are the electric field amplitude and the carrier density of the PICs. The CARP light is shown in Eq. (4) [7]. \(\varphi(t)\) is generated by the stochastic differential equation based on the Ornstein-Uhlenbeck process. The fixed parameters are \(r, G_N, N_0, \epsilon, \tau_p, \tau_s,\) and \(\alpha\). They correspond to the gain coefficient, the carrier density at transparency, the gain saturation coefficient, the photon lifetime, the carrier lifetime, and the linewidth enhancement factor, respectively. The variable parameters are \(\kappa, J, \varphi, \Delta \omega,\) and \(\Delta \omega_c\). They correspond to the feedback strength, the injection current, the feedback delay time, the injection strength, the propagation delay time from the drive laser to the PICs, the optical frequency, and the optical frequency detuning between the drive laser and the PICs, respectively. In this study, we set the parameter values as follows: \(\kappa = 10.9 \text{ ns}^{-1}, \ J_r = 1.16 \text{ J}_\text{th}, \ \tau = 0.21 \text{ ns}, \) and \(\sigma = 2.3 \text{ ns}^{-1}\). \(\Delta \omega\) describes \(\Delta \omega = 2\pi \Delta f_{\text{sol}},\) where \(\Delta f_{\text{sol}}\) is the optical frequency detuning between Drive laser and PICs \(\Delta f_{\text{sol}} = -4.0 \text{ GHz}\).

### 5. Numerical results

Figure 5 shows the temporal waveforms and the cross-correlation plots obtained from the numerical simulations. Fig.5(a) and (b) indicates high-quality synchronization of the two PIC outputs when the feedback phases of the two PICs are matched. The cross-correlation value is 1.0 in Fig. 5(b). We thus numerically achieve common-signal-induced synchronization between the PICs with CARP light. In Figs. 5(c) and (d), the feedback phase of PIC 1 is shifted to 0.6 \(\pi\). This parameter change affects the degradation of synchronization, and the cross-correlation value of two PICs is 0.038. These results are in good agreement with the experimental results in Fig. 3.

Figure 6 shows the numerical results of the fast Fourier transforms (FFT) of the temporal waveforms of the PIC outputs, corresponding to the RF spectrum of Fig. 4. In Fig. 6(a), both of the feedback phases of PICs are set to 0, and two peaks are observed and well matched in the FFTs as in the case of the experiment. On the contrary, when the feedback phase of PIC 1 is set to 0.6 \(\pi\) in Fig. 6(b), the peaks of the FFT of PIC 1 are shifted and mismatched to those of PIC 2. These results also agree with the experimental results in Fig. 4.

### 6. Conclusions

In conclusion, we experimentally achieved common-signal-induced synchronization in photonic integrated circuits driven by a constant-amplitude random-phase light. We obtained high cross correlation when the feedback phases of the PICs are matched. We investigate RF spectra of the PICs when the feedback phases are changed. Two RF spectra of the PICs are similar to each other when the feedback phases are matched. We also numerically investigated common-signal-induced synchronization of the PICs. The numerical results are in good agreement with the experimental results.

![Figure 5: Numerical results of common-signal-induced synchronization in two PICs. (a),(b) Temporal waveforms of PIC 1 and PIC 2, and (c),(d) correlation plots. The phases of the feedback lights are matched in (a),(b) and mismatched in (c),(d).](image)

![Figure 6: Numerical results of the FFTs of the PIC outputs. The feedback phases of PICs are (a) matched, and (b) mismatched.](image)

### Acknowledgments

We acknowledge support from Grants-in-Aid for Scientific Research from Japan Society for the Promotion of Science (JSPS KAKENHI Grant Number JP24686010), and Management Expenses Grants from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

### References


Demonstration of Chaos synchronization of sampled-data coupled piecewise-linear systems

HoangDao NGUYEN† and Tadashi TSUBONE

Department of Electronic Engineering, Nagaoka University of Technology
1603-1, Kamitomioka, Nagaoka, Niigata, 940-2188, Japan
†Email: s133112@stn.nagaokaut.ac.jp

Abstract—Studies on synchronization phenomena occurring in coupled oscillators have been energetically carried out from academic and engineering viewpoints. Since one of simple examples of the oscillators can be considered as one-dimensional (1-D) discrete-time dynamical systems, many approaches using 1-D chaotic maps which are equivalent to those one-dimensional systems have been studied [1]. However, considerations about the relationship between continuous-time and discrete-time systems are insufficient. In this study, we realize coupled oscillators which completely correspond to those chaotic maps. By using a simple chaos generating circuit whose Poincare-map is a 1-D chaotic map, we confirm the complete synchronization and its stability of the coupled systems. Some results are verified by laboratory measurements.

1. Introduction

A variety of nonlinear phenomena are obtained in high-dimensional systems such as coupled systems which are composed by some coupling of oscillators. For instance, synchronization, the behavior of each oscillator coincides with each other at any time, is one of the typical phenomena [1]. Chaos synchronization is a lot of research, mainly due to its potential application in secure communications. In reference [2], Kinzel et al. used a coupled system of Bernoulli maps to propose the secure communication protocol focusing on passive attackers who can only record the transmitted signals of the communication of two partners. In addition, another application is in the bistatic radar, synchronization of Lorenz systems is proposed to synchronizing the response oscillator of the radar receiver with the drive oscillator at the radar transmitter, which has been noticed as the most challenge in the setup of bistatic radar [3]. As one of the effective examples to investigate the mechanism of synchronization phenomena, many approaches of coupled systems in which one-dimensional maps are considered as oscillators have been proposed [1, 4].

In recently, a lot of researches of coupled oscillators containing time-delay have been carried out [5-7]. In case of a continuous-time system of coupled oscillators, when transmission delay is considered in coupling, the system becomes very difficult to analyze because its dimension is infinite. On the other hand, in case of a discrete-time systems of a coupled of oscillators, even though time-delay is considered in coupling, system dimension is finite so that it can be analyzed comparatively easily [8]. In practical engineering problems, because the influence of the time-delay cannot be ignored, discussion of such a discrete time system is extremely essential.

Among nonlinear phenomena observed in chaotic dynamical systems with time-delay coupling, amplitude death phenomena, which is generated when coupled oscillators drive each other to fixed points and stop oscillating, is attracting a lot of research [9-11]. However, since discrete-time systems reported in previous researches are almost mathematical models, considerations about the corresponding physical systems represented by a continuous-time dynamical systems are insufficient. Therefore, in this study, by proposing a coupled oscillators exactly corresponding to a discrete-time system and discussing its synchronization phenomena, we consider the relation of coupled oscillators represented by coupled chaotic maps and a continuous-time dynamical system represented by coupled oscillators.

In this study, the manifold piecewise-linear system (MPL) given in the reference [12] is considered as a chaotic oscillator. This system has been shown that it strictly corresponding to any one-dimensional map. Using this system, recent researches of the bifurcation phenomena with time-delay [13], occurrence of super-expanding chaos [14] and many of engineering applications such as matched filter [15], chaos radar [16], etc. have been proposed.

In this report, firstly, we propose a control method in order to synchronize phases of the solution exhibited in two manifold piecewise-linear oscillators. By this control method, we can make correspondent of the solutions of two autonomous oscillators and two chaotic maps that can be considered in the same time axis. Then, we propose a controlled system coupled by the sample values, and show that this system completely corresponds to the coupled chaotic maps. As a specific example, we construct a coupled oscillator corresponding to the coupled system of Bernoulli shift maps and tent maps, and also investigate its synchronization phenomena by using well-known method for synchronizing two identical chaotic maps [1]. In a theoretical perspective, we assume that two oscillators have same pa-
rameter. However, in practical applications, a slight different parameter between two oscillators cannot be ignored. Therefore, when we synchronize coupled oscillators, there are some differences of Bernoulli shift maps and tent maps result which will be explained in detailed in Sec. 5. Finally, some simulation results are verified by laboratory measurements.

2. Manifold piecewise linear system (MPL)

Firstly, we consider MPL in reference [12]. This system has been shown that it strictly corresponding to Bernoulli shift map. Its dynamics is described by the following ordinary differential equation with switching conditions.

\[ \dot{x} - 2\delta \dot{x} + x = \begin{cases} 1 & \text{for } x_n \geq 0 \\ -1 & \text{for } x_n < 0. \end{cases} \quad (1) \]

\( x_n \) is a dependent variable of \( x \), sampled at \( \dot{x} = 0 \) and held until the next times of \( \dot{x} = 0 \). In other words, \( h \) is a state variable corresponding to the equilibrium point in piecewise-linear regions, switches from \(-1\) to \(+1\) if \( x \geq 0 \) when \( \dot{x} = 0 \), and switches from \(+1\) to \(-1\) if \( x < 0 \) when \( \dot{x} = 0 \). Block diagram is shown in Fig. 1.

![Manifold piecewise linear system](image)

Figure 1: Manifold piecewise linear system

Sample and hold circuit (\( S/H \)) samples \( x \) when \( \dot{x} = 0 \) and holds until the next times of \( \dot{x} = 0 \). By comparing output \( x_n \) switching of \( h \) is realized. Chaos attractors of simulation and experimental result, which are shown on projected figure of plane \( (x, \dot{x}) \) (see Fig. 2), are matched qualitatively.

![Chaos attractor of MPL](image)

Figure 2: Chaos attractor of MPL

3. Poincare map

An example of solution trajectory on phase space is shown in Fig. 3. Assuming that, when \( h = 1 \) solution starts from the initial state \( x(0) \) on \( x \)-axis then solution rotates extensively, and \( \dot{x} \) becomes 0 every times \( \tau = n\pi/\omega \) \((n = 1, 2, 3,...)\), where \( \omega = \sqrt{1 - \delta^2} \). When \( x(\pi/\omega) \) go to be negative equilibrium point is switched to \( h = -1 \), and then in the same way \( \dot{x} = 0 \) every times \( \tau = n\pi/\omega \) \((n = 1, 2, 3,...)\). Thus, by considering set \( S \) of \( x \) when \( \dot{x} = 0 \) as a domain, return map \( F \) from \( S \) to \( S \) can be defined as follow:

\[ S = \{ (x, \dot{x}, h) | \dot{x} = 0 \} , \quad F : S \rightarrow S , \quad x_n \mapsto x_{n+1} \]

Note that \( x \)-coordinate of points on \( S \) corresponds to \( x_n \) defined previously. Therefore, return map \( F \) with \( x_n \) considered as a variable is described as follows:

\[ x_{n+1} = F(x_n) = \begin{cases} -A(x_n - h) + h & \text{for } x_n \geq 0 \\ -A(x_n - h) - h & \text{for } x_n < 0, \end{cases} \quad (2) \]

where, \( A = e^{\delta \pi/\omega} \).

![Example of solution trajectory on phase space](image)

Figure 3: Examples of solution trajectory on phase space

For generalization, the equilibrium point \( h \) of the system is rewriten to \( f(x_n) \). In this case, return map \( F \) is described by Eq. 3.

\[ x_{n+1} = F(x_n) = -A(x_n - f(x_n)) + f(x_n). \quad (3) \]

Therefore, MPL corresponding to desired return map \( F(x) \) can be synthesized by deriving function \( f(x) \) as follows:

\[ f(x_n) = \frac{F(x_n) + Ax_n}{A + 1}. \quad (4) \]
From here, considering the case when $F(x_n)$ is a tent map described by Eq. 5 as an example.

$$F(x_n) = \begin{cases} -A x_n + A + 1 & \text{for } x_n \geq 0 \\ k A x_n + A + 1 & \text{for } x_n < 0, \end{cases}$$

where $k > 0$ is coefficient that represents slope of $F(x_n)$ in region $x_n < 0$. The equilibrium point $f(x_n)$ of MPL corresponding to Eq. 5 is obtained as Eq. 6.

$$f(x_n) = \begin{cases} 1 & \text{for } x_n \geq 0 \\ \frac{1}{A + 1} x_n + 1 & \text{for } x_n < 0, \end{cases}$$

where $l = k + 1 > 1$ is coefficient that represents slope of $f(x_n)$ in region $x_n < 0$. Simulation and experimental results of attractor and return map when $k = 1$ are described in Fig. 5 and 6.

![Figure 5: Attractor and corresponding return map](image)

![Figure 6: Experimental results of Fig. 5](image)

5. Complete synchronization by sampled-data coupling

We consider the sampled-data coupling of two oscillators, in which time of $\dot{x}_m = 0$ and time of $\dot{x}_s = 0$ are coincide by phase synchronization method of Sec. 4. We also show that this coupled system corresponds to the coupled chaotic map.

The mutual coupled chaotic maps with slight different parameters between two systems are considered as bellow:

$$\begin{align*}
O_1 : & \quad x_{m(n+1)} = F(x_{m(n)}) - K_a(x_{2(n)} - x_{m(n)}) \\
O_2 : & \quad x_{2(n+1)} = F(x_{2(n)}) + K_a(x_{2(n)} - x_{m(n)}),
\end{align*}$$

where $F$ is considered in two cases, Bernoulli shift map and tent map, and $K_a$ is a coupling constant. Fig. 8 shows plots of $x_{O1}$ and $x_{O2}$ in two cases after a sufficient time. In case of Bernoulli shift map, it is weak synchronization. In other hand, in case of tent map, strong synchronization is generated.

Now, we consider coupled oscillators corresponding to the coupled maps.

![Figure 8: Synchronization of coupled maps](image)
\[
\begin{align*}
O_1 : & \ddot{x}_{O1} - 2\delta \dot{x}_{O1} + \omega_{O1} x_{O1} = f(x_{O1}) - \Delta h \\
O_2 : & \ddot{x}_{O2} - 2\delta \dot{x}_{O2} + (\omega_{O2} + \Delta \omega) x_{O2} = f(x_{O2}) + \Delta h.
\end{align*}
\tag{10}
\]

Control amount \(\Delta h\) is given as follows:

\[
\Delta h = K_b (x_{O2(0)} - x_{O1(0)}), \quad K_b = \frac{K_a}{1 + A}. \tag{11}
\]

Results of this system shown in Fig. 9, 10 and 11 are similar to results of coupled map (9), in the case of Bernoulli shift map, non-synchronized regions appear when two oscillators fall into 0, which is the non-continuous point of the map, and go out with reverse direction, it is called weak synchronization (see Fig. 9). In the case of tent map, synchronization is strong at \(k \simeq 1\) (see Fig. 10), and becomes weaker when coefficient \(k\) representing slope of tent map is increased (see Fig. 11).

![Figure 9: Synchronizing time waveform and its trajectory (case of Bernoulli shift map, \(\delta \simeq 0.12\)](image)

![Figure 10: Synchronizing time waveform and its trajectory (case of tent map, \(k \simeq 1, \delta \simeq 0.12\)](image)

6. Conclusion

In this report, firstly, a control method was proposed in order to synchronize phases of the solution exhibited in two manifold piecewise linear system oscillators. By this control method, solutions of two autonomous oscillators correspond to two chaotic maps that are considered in the same time axis. Then, a coupled system coupled by sampled-data was proposed, and shown that this system completely corresponds to the coupled chaotic map. As a specific example, a coupled oscillators system corresponding to the coupled system of well-known Bernoulli shift maps and tent maps was constructed, and its synchronization phenomena was also investigated.

References

Investigation of Phase Itinerancy of Complex Waves on a Ring Constructed by Van Der Pol Oscillators

Ryouhei Takano† and Masayuki Yamauchi‡

†Hiroshima Institute of Thechnology
2-1-1 Miyake, Saeki-ku, Hiroshima, 731-5193 Japan
Email: a113115@cc.it-hiroshima.ac.jp, yamauchi@ieee.org

Abstract—In our previous study, we analyzed synchronization phenomena and confirmed the waves which intricately behave on van der Pol oscillators coupled by inductors as a ring. In this study, we investigate the details of this complex waves by using phase states, phase differences between adjacent oscillators and instantaneous electric powers in changing the number of van der Pol oscillators.

1. Introduction

In this world, there is a synchronization phenomenon which is one of important phenomena. Synchronization phenomena are shown as flashing groups of fireflies in south-east Asian, as motion among pacemaker cells, as a relationship of the rotation and the revolution of the moon, as a generating phenomena of laser on semiconductor, and so on[1]-[4]. Especially, the synchronization phenomena can be easily observed as one of quickly and clearly phenomena in electric circuits.

In our previous study, we investigated and analyzed synchronization phenomena on van der Pol oscillators coupled by inductors as a ring. We discovered and observed continuously propagating wave motions with switching phase states between adjacent oscillators. The wave motions are named phase-inversion waves[5]. We also observed special waves which propagate a phase difference between adjacent oscillators and change to the phase-inversion waves or disappear. The waves are called phase-waves[6]. Furthermore, complex waves, which are not phase-inversion waves and phase-waves and continuously propagate, were discovered. The complex wave can be classified to two types. One of the complex waves is waves like mixed phase-inversion waves and phase-waves, and the other one is winding waves.

In this paper, we investigate these complex waves. Relationships between the complex waves and itineraries of phase states are analyzed, and the details of complex waves are investigated by using phase differences between adjacent oscillators, instantaneous electric powers, and so on, in changing the number of oscillators.

2. Circuit model

Our circuit model is shown in Fig. 1. N van der Pol oscillators are coupled by inductors \( L_c \) as a ring. Each van der Pol oscillator is constructed by using a inductor \( L \), a capacitor \( C \) and a nonlinear negative resistor \( f(\nu) \). The \( f(\nu_k) \) of \( k \)-th oscillator(Oscillator \( k \)) is assumed as Eq. (1). The Oscillator \( k \) is written as OSC\(_k\) in this paper.

\[
f(\nu_k) = -g_1\nu_k + g_3\nu_k^3 \quad (1 \leq k \leq N)
\]

Circuit equations of this circuit are normalized by using Eq. (2). The normalized equations are shown in Eq. (3).

\[
i_k = \sqrt{\frac{Cg_1}{3Lg_3}}x_k, \quad v_k = \sqrt{\frac{g_1}{3g_3}}y_k, \quad t = \sqrt{\frac{LC}{\epsilon}},
\]

\[
\alpha = \frac{L}{L_c}, \quad \epsilon = \frac{g_1}{3}, \quad \delta = \frac{g_1^2}{3g_3}.
\]

\[
\frac{dx_k}{dt} = y_k,
\]

\[
\frac{dy_k}{dt} = -x_k + \alpha (x_a - 2x_k + x_b) + \epsilon \left(y_k - \frac{1}{3}y_k^3\right).
\]

(If \( k = 1, \quad \alpha = N \) and \( b = 2 \). If \( k = N, \quad \alpha = N - 1 \) and \( b = 1 \). If \( 2 \leq k \leq N - 1, \quad \alpha = k - 1 \) and \( b = k + 1 \).)

The instantaneous electric powers are calculated by using Eqs. (4)–(5). The \( P_k \) shows an instantaneous electric power of OSC\(_k\). The \( P_L(k-1,k) \) shows an instantaneous electric power of coupling inductor \( L_c \) between OSC\(_{k-1}\) and OSC\(_k\).

\[
P_k = \frac{\alpha\delta}{\epsilon} y_k (x_a - 2x_k + x_b)
\]

\[
P_L(k,a,k) = \frac{\alpha\delta}{\epsilon} (x_k - x_a)(y_k - y_a)
\]

(If \( k = 1, \quad \alpha = N \) and \( b = 2 \). If \( k = N, \quad \alpha = N - 1 \) and \( b = 1 \). If \( 2 \leq k \leq N - 1, \quad \alpha = k - 1 \) and \( b = k + 1 \).)

Normalized circuit equations of this circuit model are simulated by using fourth order Runge-Kutta method.

3. Phase Itinerancy of Complex Waves

Complex waves are investigated on the ring. We set 5 observation conditions as follows.
1. N is changed from 9 to 16.
2. δ is fixed as 1.
3. α is fixed as 0.50.
4. ε is fixed as 0.35.
5. The phase-inversion waves are generated in the in-phase synchronizations and initial values which a set of two phase-inversion waves propagates are set.

Some differences between phase differences itinerancies of two types of complex waves are investigated.

3.1. Changing the number of oscillator

The itinerancies of phase differences are shown in Figs. 2–7. The Figs. 2–4 are constructed by stacking long and thin boxes. In each box, sum of voltages of adjacent oscillators is shown along time. A phase state between OSC1 and OSC2 is shown in the top box, and a phase state between OSCN and OSC1 is shown in the bottom box. Therefore, black regions express the almost in-phase synchronization and white regions express the almost anti-phase synchronization. In the Fig. 2, we can confirm many waves propagate and disappear. The winding complex waves can be observed in the Figs. 3 and 4. When the number of oscillator is increased, the width of black regions and white regions are increased too (see Figs. 3–4). The Figs. 5–7 show itinerancies of phase differences of adjacent oscillators. When the Figs. 2–4 are observed, we can understand that itinerancies are very complex.

3.2. Two types of complex waves

The complex waves can be classified to two types.

Type A The waves look like mixed phase-inversion waves and phase-waves.

Type B The waves is winding and propagating.

The types of the complex waves are shown in Table 1 when the number of oscillators is changed from 9 to 16. We investigate relationships between itinerancies of phase differences and the complex wave of the each type. Itineraries of phase differences of the Type A and Type B are shown in Figs. 8 and 9, respectively. We can observe phase differences which are continuously expanding along time in the Fig. 8 when the Type A complex waves are propagating. However, in the Fig. 9, we can confirm phase differences which do not continuously expands when the complex waves of Type B are propagating. Characteristics of Type A complex waves differ from characteristics of Type B complex waves.
Figure 4: Winding complex waves($N = 16$, $\alpha = 0.50$, and $\epsilon = 0.35$)

Figure 5: Itinerancies of phase differences of complex waves($N = 9$, $\alpha = 0.50$, and $\epsilon = 0.35$)

3.3. Detail investigation by using instantaneous electric powers

The relationship between complex waves and instantaneous electric powers are investigated. The instantaneous electric powers are shown in Figs. 10–12. The Fig. 10 shows instantaneous electric powers of OSC$_2$ and OSC$_3$ of the Type A complex wave on the nine oscillators array, and the Figs. 11 and 12 show instantaneous electric powers of OSC$_2$ and OSC$_3$ of the Type B complex wave on the 10 oscillators array and the 16 oscillators array. The itinerancy of the instantaneous electric power of the Type A complex wave differ from the itinerancy of the instantaneous electric power of the Type B. The itinerancy of instantaneous electric power of the Type A is an irregular pattern and very complex (see the Fig. 10). However, the itinerancy of the Type B is not the irregular pattern (see the Figs. 11 and 12). Characteristics of Type A complex waves differ from characteristics of Type B complex waves.

In the Figs. 11 and 12, a period of the large amplitude shows the white regions of the Figs. 3 and 4, and a period of the small amplitude shows the black regions of the Figs. 3 and 4. When phase-inversion waves are propagating, the widths of white regions are fixed. However, when complex waves are propagating, the width of white regions are fluctuated. We can understand that the rate of black and white regions are almost not different when the number of oscillator is changed (see the Figs. 11 and 12).

4. Conclusions

We made clear that complex waves are effected by the number of oscillators and the complex waves can be classified to two types. When the Type A complex waves were propagating, the phase differences of each oscillators expanded along time. However, we made clear that the phase differences of each oscillators do not expand when the Type B complex waves propagate. Moreover, an itinerancy instantaneous electric power of Type A differed from an itinerancy instantaneous electric power of Type B. The itinerancies of instantaneous electric powers of Type B was almost regular, but itinerancies of instantaneous electric powers of Type A was not regular. We clarified that character-

<table>
<thead>
<tr>
<th>Table 1: Type of complex waves</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>$A$</td>
</tr>
</tbody>
</table>
istics of Type A complex waves differ from characteristics of Type B complex waves. Furthermore, we observed that the rate of black and white regions are almost not different when the number of oscillators is changed and Type B complex waves can be observed.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 25730152.

References


Synchronization in Dynamical Polygonal Oscillatory Networks with Switching Topology

Yoko Uwate†, Yoshifumi Nishio† and Ruedi Stoop‡

†Dept. of Electrical and Electronic Engineering, Tokushima University
2-1 Minami-Josanjima, Tokushima, Japan
Email: {uwate, nishio}@ee.tokushima-u.ac.jp
‡Institute of Neuroinformatics, University/ETH Zurich,
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
Email: ruedi@ini.phys.ethz.ch

Abstract— In this study, we propose dynamical polygonal oscillator networks with switching topology by using van der Pol oscillators. In the proposed network, the network topology is switching with time. We confirmed that new synchronization state can be occurred by stochastically switching the network topology.

1. Introduction

The synchronization phenomena observed from coupled nonlinear oscillators are suitable model to analyze the natural phenomena [1],[2]. Therefore, many researchers have proposed different coupled oscillatory networks and have discovered many interesting synchronization phenomena [3]-[6].

In the human brain, synchronization of neural activity between different cortical areas may proved several kinds of neural information functions depending on the network topology (see Fig. 1(a)) [7]. Recently, grid cells which are the type of neurons in the entorhinal cortex show remarkable hexagonal activity patterns [8], [9]. The polygonal structure of neuronal firing pattern has possibility important role for emerging unified computational framework [10].

Namely, it is important to study synchronization phenomena observed from the polygonal oscillatory networks with the several characteristics in the human brain as we described above, for modeling neuro-biological systems and applying its high-functional information processing to engineering applications.

Neurons in the brain are expressed by stable oscillators. Neuronal oscillators using mathematical models are often used for modeling brain networks. However, there are not many models of neuronal oscillators using electrical circuits. In order to realize the brain network by real physical systems, we need to investigate synchronization of neural activity in analog electrical circuits. Here, van der Pol oscillator which is simple oscillatory circuit model is used as neural oscillator.

We have investigated the synchronization phenomena in the coupled polygonal oscillatory networks sharing branches [11], [12]. In this system, odd number of van der Pol oscillators are connected to every corner of each polygonal network. Namely, frustration is occurred between the adjacent oscillators. By using computer simulations and theoretical analysis, we confirm that the coupled oscillators tend to synchronize to minimize the power consumption of the whole system. The phase difference of the shared oscillators is solved by finding the minimum value of the power consumption function.

Furthermore, coupling strength between neural activities are changed depending on the information processing with time (see Fig. 1(b)). However, in Ref. [11], we have considered the static network model.

In this study, we propose new dynamical polygonal circuit system which is including the switching couplings, in order to understand the mechanism of high functional neural information processing in the human brain. In the proposed network, external signal is replaced with stochastic factor and we consider that the coupling strength is expressed by on/off as an extreme example. The synchronization phenomena in the proposed networks are investigated by using computer simulations. First, we investigate synchronization states in two coupled polygonal networks (triangular and quadrangular networks) with one switching coupling, in order to understand the basic phenomena. Next, the system model is extended to six coupled polygonal networks with three switching couplings for understanding more complex behavior. By using the computer simulations, we confirm that new synchronization states
can be produced when the network topology is changed by switching the connection (on/off) of the edges in the proposed network.

2. Two Coupled Network
- Number of Switching Coupling : 1 -

2.1. Circuit Model

Figures 2 and 4 show the conceptual circuit model and circuit realization using van der Pol oscillators. The triangular and the quadrangular oscillatory networks are coupled by sharing a branch. We call this circuit system “3-4 coupling network.” In this model, we consider the coupling method which two adjacent oscillators are tend to synchronize at anti-phase state.

In Fig. 2, a chain line denotes the switching coupling (SC). If the coupling edge is selected as the switching coupling, the two oscillators connect or disconnect according to the coupling probability ($p$). Namely, the switching coupling is connecting and disconnecting stochastically. If the coupling probability is $p = 0.0$, the coupling strength is $\gamma = 0$. The state (connecting/disconnecting) of the switching coupling is updated at every switching time (ST) in the simulation. The switching time denotes the iteration number. One example of operation of the switching coupling is shown in Fig. 3.

![Figure 2: Conceptual circuit model for 3-4 coupling network.](image)

We consider only one coupling edge can be the switching coupling in the 3-4 coupling network. Table I is summarized the network typologies and synchronization states of four patterns when the switching coupling is connecting and disconnecting. We consider the four patterns (pattern-A, B, C, and D) depending on the switching coupling. “state-1” and “state-2” denote the synchronization state obtained from the two types of the network topologies.

![Figure 3: Example of operation of switching coupling for 3-4 coupling network.](image)

Table 1: Switching Coupling (SC) Patterns.

<table>
<thead>
<tr>
<th>Pattern</th>
<th>state-1 (SC: on)</th>
<th>state-2 (SC: off)</th>
</tr>
</thead>
<tbody>
<tr>
<td>pattern-A (SC: $\gamma_{12}$)</td>
<td>$\phi_{12} \approx 138.6^\circ$</td>
<td>$\phi_{12} \approx 72^\circ$</td>
</tr>
<tr>
<td>pattern-B (SC: $\gamma_{13}$)</td>
<td>$\phi_{12} \approx 138.6^\circ$</td>
<td>$\phi_{12} \approx 180^\circ$</td>
</tr>
<tr>
<td>pattern-C (SC: $\gamma_{14}$)</td>
<td>$\phi_{12} \approx 138.6^\circ$</td>
<td>$\phi_{12} \approx 120^\circ$</td>
</tr>
<tr>
<td>pattern-D (SC: $\gamma_{15}$)</td>
<td>$\phi_{12} \approx 138.6^\circ$</td>
<td>$\phi_{12} \approx 120^\circ$</td>
</tr>
</tbody>
</table>

![Figure 4: Circuit model for 3-4 coupling network.](image)

Next, we develop the expression for the circuit equations of 3-4 coupling oscillatory networks as shown in Fig. 4. The $v_k - i_{rk}$ characteristics of the nonlinear resistor are approximated by the following third order polynomial equation,

$$i_{rk} = -g_1v_k + g_3v_k^3 \quad (g_1, g_3 > 0), \quad (k = 1, 2, 3, 4). \quad (1)$$

The normalized circuit equations governing the circuit are expressed as [4th oscillator]

$$\begin{align*}
\frac{dx_k}{d\tau} &= \epsilon \left( 1 - \frac{1}{3}v_k^2 \right) x_k - (y_{ak} + y_{bk} + y_{ck}) \\
\frac{dy_{ak}}{d\tau} &= \frac{1}{3} \left( x_k - \eta y_{ak} - \gamma_{kn}(y_{ak} + y_n) \right) \\
\frac{dy_{bk}}{d\tau} &= \frac{1}{3} \left( x_k - \eta y_{bk} - \gamma_{kn}(y_{bk} + y_n) \right) \\
\frac{dy_{ck}}{d\tau} &= \frac{1}{3} \left( x_k - \eta y_{ck} - \gamma_{kn}(y_{ck} + y_n) \right) \\
\end{align*}$$

(2)
where
\[ t = \sqrt{LC} \tau, \quad v_k = \sqrt{\frac{g_1}{3g_3}x_k}, \quad i_{ak} = \sqrt{\frac{g_1}{3g_3}x_k}, \]
\[ i_{bk} = \sqrt{\frac{g_1}{3g_3}} \sqrt{\frac{C}{L}} y_{bk}, \quad \varepsilon = g_1 \sqrt{\frac{L}{C}}, \]
\[ \gamma = R \sqrt{\frac{C}{L}}, \quad \eta = r_m \sqrt{\frac{C}{L}}, \]
\[ (k = 1, 2, 3, 4, 5). \]

In these equations, \( x_{kn} \) is the coupling strength, \( \varepsilon \) denotes the nonlinearity of the oscillators and \( v_k \) denotes the current of neighbor oscillator on the coupling resistor. For the computer simulations, we calculate Eq. (2) using the fourth-order Runge-Kutta method with the step size \( h = 0.005 \). The parameters of this circuit model are fixed as \( \varepsilon = 0.1 \) and \( \eta = 0.0001 \).

2.2. Synchronization Phenomena

First, we investigate the synchronization state of the 3-4 coupling network with the switching topology. The switching time (ST) is changed from 100 to 10000. Figures 5 and 6 show the examples of the phase difference between 1st and 2nd oscillators for Patterns A and C at Table 1 when the coupling probability is set to \( p = 0.5 \). The phase difference is calculated by using the Poincaré section: \( x_k < 0, y_k = 0 \).

Figure 5: Example of the phase difference for Pattern A \((p=0.5)\).

Figs. 5(d) and 6(d)). And the converging phase state shows value between the state-1 and the state-2.

Figure 7 shows the simulation results of observing new phase states by changing the coupling probability \( p \) for four patterns. The phase difference shows the average value of 1000 iterations in the simulation.

In the cases of Patterns-A, C and D, the new phase states can be observed between state-1 and state-2 in the range of \( 0.1 \leq p \leq 0.9 \). While, in the case of Pattern-B, the phase states show almost similar value with state-2 when the coupling probability is smaller than 0.6. By increasing \( p \), the new phase states can be obtained as shown in Fig. 7(b).

From these results, we can see that new synchronization states can be occurred by changing the network topology.

3. Conclusions

In this study, we have proposed new dynamical polygonal circuit system which is including the switching couplings, in order to understand the mechanism of high functional neural information processing in the human brain. In the proposed network, external signal is replaced with stochastic factor and we consider that the coupling strength is expressed by on/off as an extreme example. By using the computer simulations, we confirmed that new synchronization state can be occurred by switching the network topology.

For the future works, we investigate the synchronization state in detail when the three switching couplings is changed with different iteration in the extended dynamical network. And, we would like to apply the proposed system to more large scale networks to model the existing biological complex networks.
Figure 7: Observing new phase states by changing $p$.

References


Review of Tuned Power Oscillators

Hiroo Sekiya†

†Graduate School of Advanced Integration Sciences, Chiba University
1-33, Yayoi-cho, Inage-ku, Chiba, 263-8522 Japan
Email: sekiya@faculty.chiba-u.jp

Abstract—This paper presents a review of tuned power oscillators. Tuned power oscillator provides ac output power from dc input power by using switching devices. One of the most important points of tuned power oscillator is that the switching devices are driven by the feedback voltage from the ac output. It is possible to make a self-oscillation by using the feedback mechanism. This paper introduces some tuned power oscillator for high-frequency applications. Each oscillator achieves high power-conversion efficiency by applying the soft switching technique.

1. Introduction

The increase in the power density is a major purpose in the power electronics research field. For enhancing the power density, the circuit scale should be smaller. The magnetic components such as inductor and transformer are dominant factor for determining the circuit volume. Therefore, it can be stated that high frequency operation is effective to reduce the circuit volume. Switching-device driver design is a covert problem in high-speed switching devices. The square-driving voltage is strained due to parasitic capacitances and resistances of the switching device. Because a precise switching pattern is required for achieving the soft-switching conditions at high frequencies, the driver-circuit designs, taking into account the parasitic components, are the technical barrier for high-frequency power-electronics circuit.

The tuned power oscillators, which are autonomous circuit without driver, is one of the solutions for high-frequency power-electronics circuit. The tuned power oscillator is driven by the feedback signal from the sinusoidal output voltage. In addition, it is possible to apply the soft-switching techniques and high power-conversion efficiency can be achieved. Namely, the tuned power oscillators are suitable to high-power density converters. Actually, there are wide-area applications of tuned power oscillators, for example dc-ac inverter part of the dc-dc converters and transmitters of the wireless power transfer systems and wireless communications. Tuned power oscillators, however, have problems of design difficulty and frequency unstability. It is useful and effective for tuned-power oscillator usages to understand the operation mechanism and design strategies of high-frequency high-efficiency tuned power oscillators.

2. Review of Tuned Power Oscillator

A fundamental configuration diagram of tuned power oscillators is shown in Fig. 1. The switching devices of tuned power oscillators are driven the voltage through the feedback network from the output voltage. By applying the feedback voltage as the driving signal, the circuit works with self-oscillation and designers can be relieved from the implementation difficulty of driver circuit. The driving signal is not a square waveform but a sinusoidal waveform in the tuned power oscillator because the output voltage is regarded as a sinusoidal waveform. Namely, the feedback network should have roles to adjust phase shift between output voltage and gate-source voltage and amplitude of the gate voltage, which should be less than the permissive value.
2.1. Free-Running Class-E Oscillator

Figure 2(a) shows a circuit topology of the free-running class-E oscillator [7]-[11]. The class-E oscillator consists of the class-E amplifier and feedback network $C_1$, $C_2$ and $L_f$. $R_{d1}$ and $R_{d2}$ give the bias voltage, which is the same as the threshold voltage $V_{th}$, for the gate of the switching device. Figure 3 shows example waveforms of the class-E oscillator with nominal conditions. The switching device of the class-E oscillator is driven by the feedback voltage $v_f$, which is from the output voltage $v_o$. The feedback voltage is a sinusoidal waveform because the feedback current flow through the resonant filter, which consists of $L_f$, gate-source parasitic capacitance $C_g$, and gate-source parasitic resistance $r_g$. Because $C_g$ and $r_g$ are fixed, which depend on the MOSFET type, the feedback network can be designed by choosing the component values of $C_1$, $C_2$, and $L_f$. By adjusting them, the amplitude and phase shift between the output voltage and the gate voltage are adjusted. The fundamental operation is the same as the class-E amplifier [12]-[15]. Namely, the switch voltage achieves the class-E zero-voltage switching and zero-voltage-derivative switching (ZVS/ZVDS) conditions. Because of the class-E ZVS/ZDS conditions, the class-E oscillator achieves high power-conversion efficiency at high frequencies.

2.2. Injection-locked Class-E Oscillator

Figure 5 shows an example topology of the injection-locked class-E oscillator [16], [17]. The small-power signal $v_{inj}$ is injected to the gate terminal as shown in Fig. 5. Because the injection-signal power is low, it is possible to obtain the injection-locked oscillator by just adding the injection signal to the original free-running oscillator. If the feedback voltage of the class-E free-running oscillator is synchronized with the injection signal $v_{inj}$, the oscillator frequency is locked with the injection-signal frequency $f_{inj}$, which means the frequency of the output voltage is fixed with $f_{inj}$. It is easy to achieve synchronization as the injection-signal power increases. However, high power injection affects the waveforms of the feedback voltage and switch-on duty ratio, which yields the design complexity. It is necessary to conduct the total design of the free-running oscillator and injection circuits for large perturbation. Additionally, low injection-signal power is good from a power-added efficiency perspective.

2.3. Class-E_M Oscillator With Second Harmonic Injection

Figure 5(a) shows a circuit topology of the class-E oscillator with second harmonic injection [18], which is composed of the main circuit and the injection circuit. The injection circuit is usually operated as the class-E frequency doubler [19], [20]. The nominal waveforms of the oscillator are shown in Fig. 5(b). The main circuit is driven by the feedback voltage $v_f$ from the output voltage. The switch voltage $v_{S1}$ satisfies the class-E ZVS/ZVDS conditions at transistor turn-on instant. Additionally, the switch current $i_{S1}$ achieves the zero-current switching (ZCS) and zero-current-derivative switching (ZCDS) conditions simultaneously at the transistor turn-off instant. Because of the ZCS/ZCDS conditions, the waveforms of both the switch voltage and current at the transistor turn-off are also smooth. Because of these switching conditions, which are called the class-E_M ZVS/ZVDS/ZCS/ZCDS conditions, there are no jumps on the switch-voltage and switch-current waveforms in the main circuit. Therefore, the
class-EM amplifier enhances high power conversion efficiency even if the main-circuit transistor has long turn-off-switching time and suppresses the implementation cost. For achieving the ZCS/ZCDS conditions in the class-EM amplifier, the injection circuit is mandatory [21]-[23]. The injection circuit should provide the second-harmonic current \( i_2 \) with the proper phase-shift and the proper amplitude for achieving the ZCS/ZCDS conditions in the main-circuit switch. The switch voltage \( v_{S1} \) is transformed into the sinusoidal output voltage \( v_o \) through the resonant filter \( L_1 - C_1 \). The injection circuit is driven by the input signal \( s_m \) whose fundamental frequency is the same as the output voltage. In other words, the output frequency is locked with the input frequency. In this sense, the proposed oscillator is regarded as one of the injection-locked oscillators.

From the above explanations, it can be stated that the injection circuit has multiple roles in the class-EM oscillator. First, it offers the class-EM ZVS/ZVDS/ZCS/ZCDS conditions, which enhance the power-conversion efficiency and allow to use a slow switching device. It is possible to reduce the circuit-implementation cost, especially, the main-circuit-MOSFET cost. Second, the output-voltage frequency is locked with the input-signal frequency, which is half as high as the injection-current frequency. Finally, the output power becomes high by adding the injection circuit, which is useful for high-power applications.

3. Conclusion

This paper has presented a review of tuned power oscillators. Tuned power oscillator provides ac output power from dc input power by using switching devices. One of the most important points of tuned power oscillator is that the switching devices are driven by the feedback voltage from the ac output. It is possible to make a self-oscillation by using the feedback mechanism. This paper introduces some tuned power oscillators for high-frequency applications. Each oscillator achieves high power-conversion efficiency by applying the soft switching technique.

Acknowledgement

This research was partially supported by Grants-in-Aid for Scientific Research (No. 26289115) of JSPS, Japan.

References


Intrinsic Localized Modes in Saturable Inductor Transmission Lines

M. Sato†, T. Mukaide†, T. Nakaguchi†, 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: msato153@staff.kanazawa-u.ac.jp

Abstract—Nonlinear 1-D electric transmission lines have long been used to study solitons and intrinsic localized modes (ILMs) with the focus mainly on nonlinear capacitors since they correspond to nonlinear potential energy terms in the corresponding mechanical systems. Here, we study a saturable inductor in an otherwise linear transmission line. Our simulations show ILM current waveforms strongly distorted from a sinusoidal time dependence. The well known rotating wave approximation fails to predict such an ILM; however, including the fundamental and the third harmonic of the ILM current produces results in good agreement with simulations over a restricted amplitude region.

1. Introduction

Fundamental studies focusing on a localized nonlinear excitation with width comparable to the lattice constant of a lumped nonlinear electrical transmission line have appeared in the last decade[1-4]. To date all of these intrinsic localized mode (ILM) systems have made use of nonlinear capacitors to produce intersite nonlinear coupling between the linear inductor lattice sites. In the asymptotic strongly localized limit the excitation extends over three lattice cells.

In this report we describe a different kind of ILM associated with nonlinear inductors equally spaced in an otherwise linear electrical transmission line. In an earlier work, we considered a transmission line with a flux dependent inductance, where the rotating wave approximation (RWA) could be used. [5] Here a 1-D electric lattice with linear intersite capacitance coupling plus a current dependent inductance is the starting point. Our results demonstrate that an RWA that makes use of only the fundamental frequency is not sufficient to provide a realistic current ILM since strong harmonics appear in the equations of motion. Including these harmonic contributions shows that in the asymptotic strongly localized limit the excitation can extend over a single lattice cell.

2. Inductor model, circuit and equations of motion

A simple nonlinear inductor model with no hysteresis that describes saturation of the total flux \( \Phi \) with respect to the current \( I \) is

\[
I = \frac{1}{\sqrt{\eta}} \tan \left( \sqrt{\frac{\eta}{L_0}} \Phi \right)
\]

(1)

where \( L_0 \) is the linear inductance and \( \eta \) is the nonlinear parameter. Figure 1 illustrates this dependence of the flux on the current. Also shown is the corresponding nonlinear inductance

\[
L(I) = \frac{d\Phi}{dI} = \frac{L_0}{1 + \eta I^2}
\]

(2)

which decreases with increasing current.

![Fig. 1. (a) Current dependence of the total flux \( \Phi \) through the inductor. (b) The nonlinear inductor \( L(I) \) for the model describe by Eq. (1).](image)

The transmission line under consideration is shown in Fig. 2. Because for the simulations we are interested in exciting the plane wave zone boundary mode to produce an ILM above the plane wave spectrum the transmission line is coupled to a set of drivers with opposite phase.

In the absence of the driver and the resistance the equations of motion of the current in Fig. 2 can be written as

\[
\frac{1}{C} \int J_{n-1} dt = L(I_n) \frac{dI_n}{dt} + \frac{1}{C} \int J_n dt
\]

(3)
where the relative currents are described by 
\[ J_{n-1} = I_{n-1} - I_n \quad J_n = I_n - I_{n+1} \] and \( L \) is given by Eq. (2). 

Eq. (3) is readily transformed into 
\[
\frac{d}{dt} \left[ \frac{i_n}{1 + \eta I_n^2} \right] = -\frac{\omega_m^2}{4} \left( 2I_n - I_{n+1} - I_{n-1} \right)
\] (4)

where \( \omega_m = \sqrt{\frac{2}{L_i C}} \) is the maximum frequency of the linear dispersion curve \( 2\omega_m^2 = \omega_m^2 \left( 1 - \cos k \right) \). Equation (4) is similar to that for a spring-mass transmission line system with an amplitude dependent mass. The corresponding mass would become smaller with increasing amplitude.

3. Driven-damped simulations

Simulations using a driven damped lattice were performed to investigate the ILM properties. Damping is caused by the resistance \( R \) in Fig. 2. The circuit is driven by an oscillator via coupling capacitors \( C_d \) as shown. Because the nonlinearity in Eq. (4) is positive, we expect the generation of an ILM to commence at the zone boundary where the normal plane wave mode frequency is largest.

The equation of motion now becomes 
\[
\dot{I}_n \left( 1 + \eta I_n^2 \right) = 2\eta I_n I_n^2 \\
-\frac{\omega_m^2}{4} \left( 2I_n - I_{n+1} - I_{n-1} \right) \left( 1 + 2\eta I_n^2 + \eta^2 I_n^4 \right),
\] (5)

where last two terms are the damping and the driver, respectively.

The resulting stationary ILM is shown in Fig. 3. Figure 3(a) displays the time dependence of the ILM, and Fig. 3(b) shows its spatial pattern. The shape is appropriate for an odd symmetry mode. The driven-damped simulation of the frequency squared versus the amplitude squared is summarized by the dashed curve in Fig. 4, indicating that the mode frequency is nearly proportional to the amplitude.
4. Testing the rotating wave approximation

For the first method of analysis to compare with the simulation results we used the RWA. We assume sinusoidal time-dependence of the current,

\[ I_n = A_n \cos \omega t \]  

and then apply the RWA to Eq. (4) to obtain the following algebraic equations

\[
-\omega^2 A_n \left(1 + \frac{3}{4} \eta A_n^2\right) = \omega^2 \frac{1}{2} \eta A_n^2 A_n \\
-\frac{\omega^2}{4} \left(2A_n - A_{n+1} - A_{n-1}\right) \left(1 + \frac{3}{2} \eta A_n^2 + \frac{5}{8} \eta^2 A_n^4\right). 
\]  

This set of nonlinear equations is solved using Powell’s hybrid method in MINPACK software. The dependence of the ILM frequency on amplitude is represented by the highest frequency curve shown in Fig. 4. The squared frequency of the ILM increases linearly with \( \eta A_n^2 \) but with a larger slope than does the driven-damped simulation (dashed curve).

![Graph showing time dependence of simulated ILM current](image)

Fig. 5. Time dependence of the simulated ILM current at \( 1.25 f_m \). It is time periodic but distorted from a sinusoidal curve by higher harmonics.

The reason for the difference becomes obvious when one examines the time dependence of the ILM current as obtained from the driven-damped simulations. The time dependence of the current shown in Fig. 5 is very different from that given by Eq. (6).

The next step in analytic complexity is to add a third harmonic term to the ILM description. The new approximation becomes

\[ I_n = A_n \cos \omega t + B_n \cos 3\omega t \]  

The new coefficients for the fundamental and 3rd harmonic signals are now determined from Eq. (4) in a way similar to the RWA method. The term that invalidates the RWA is the first term of right hand side of Eq. (9):

\[
\hat{I}_n(1 + \eta I_n^2) = \frac{2\eta I_n^2}{1 + \eta I_n^2} - \frac{\omega^2}{4}(2I_n - I_{n+1} - I_{n-1})(1 + \eta I_n^2). 
\]  

If this term were absent, the usual RWA would work well. To proceed we expand the denominator in Eq. (9) up to 4th order in current so that

\[
\hat{I}_n = 2\eta I_n^2 \left(1 - \eta I_n^2 + \eta^2 I_n^4\right) \\
- \frac{\omega^2}{4} \left(2I_n - I_{n+1} - I_{n-1}\right)(1 + \eta I_n^2). 
\]

All terms in Eq. (10) are calculated using Eq. (8). Although tedious the calculation is straightforward. An example is given in Eq. (11).
\[-9\omega^2 B_n + \frac{\omega^2}{4} (2B_{n-1} - B_{n+1})\]
\[+ \omega^2 \eta \left( \frac{1}{2} A_n^3 - A_n^2 B_n - \frac{9}{2} B_n^2 \right)\]
\[+ \frac{\omega^2}{8} \eta \left( A_n^3 + 6A_n^2 B_n + 3B_n^3 \right)\]
\[- \frac{\omega^2}{16} \eta \left( A_n^3 + 4A_n B_n \right) (A_{n+1} + A_{n-1})\]
\[+ \frac{1}{8} \eta^2 \omega^3 \left( A_n^5 + 6A_n^4 B_n + 18A_n^3 B_n^2 \right)\]
\[+ \frac{60A_n^2 B_n^3 + 18B_n^7}{8}\]
\[- \frac{1}{32} \omega^3 \eta \left( -A_n^5 + 21A_n^4 B_n + 105A_n^3 B_n^2 + 330A_n^2 B_n^3 \right)\]
\[+ 250A_n B_n^4 + 380A_n^2 B_n^5 + 45B_n^7 \]
\[= 0\]

Solving the two set of equations for N lattice points, we obtain the squared frequency as a function of $\eta A^2$ as shown by the lowest frequency trace in Fig. 4. Good agreement with the simulations does occur but only below $\eta A^2 < 0.4$.

5. Discussion and Summary

The source of the periodic sharp peaking of the ILM current with time is the nonlinear inductance. That is, the inductance becomes the smallest at the current peak. Although the slope of the current is zero at the peak, it changes more rapidly near the peak than for the constant inductance case and this feature produces the narrowing of the current peak shown in Fig. 5.

Above an amplitude of $\eta A > 0.4$, the rate of change of the inductance becomes larger as shown in Fig. 1(b). This property introduces higher harmonics than used in Eq. (7). For this reason, our second attempt at an analytical determination of the ILM properties failed for large amplitudes. Thus, the general features of an ILM associated with current saturable inductors in a 1D transmission line are threefold: (1) the RWA can not be applied to obtain an analytical solution, (2) in the asymptotic limit the ILM excitation becomes localized on a single lattice cell and (3) this ILM contains many harmonics of the fundamental ILM frequency.

Acknowledgement

M. S. was supported by JSPS-Grant-in-Aid for Scientific Research No. 25400394. A. J. S. acknowledges the hospitality of the Department of Physics and Astronomy, University of Denver, where some of this work was completed.
Search for Discrete Breathers in Unstrained Graphene

Sergey V. Dmitirev†, Ivan P. Lobzenko†,‡ and Elena A. Korznikova‡

†Institute for Metals Superplasticity Problems of Russian Academy of Sciences, 39 Khalturin St., Ufa 450001, Russia
‡Institute of Molecule and Crystal Physics of Russian Academy of Sciences, 151 Prospekt Oktyabrya, Ufa 450075, Russia

Email: dmitriev.sergey.v@gmail.com, ivanlobzenko@gmail.com, elena.a.korznikova@gmail.com

Abstract – Graphene is one atom thick carbon layer with unique combination of physical and mechanical properties promising for many applications. Nonlinear lattice dynamics of graphene has been addressed in recent theoretical and experimental works. In particular, several studies have been done on the possibility to excite discrete breathers (DBs), which are long-lived, spatially localized nonlinear vibrational modes. In the present work, based on the molecular dynamics simulations, we discuss the possibility to excite DBs in unstrained graphene. It is concluded that the results are strongly dependent on the interatomic potentials used in the simulations. Our study calls for additional studies based on ab initio calculations to prove or disprove the existence of DBs in unstrained graphene.

1. Introduction

Graphene attracts attention of many researchers because it is a promising material in a number of applications ranging from electronics and solar cells to supercapacitors and hydrogen storage devices [1]. Of great interest is the nonlinear dynamics of graphene lattice, in particular, the possibility to excite spatially localized nonlinear vibrational modes, called intrinsic localized modes or discrete breathers (DB), which have been studied experimentally [2] and theoretically [3-15]. DBs in graphene and carbon nanotubes have been studied by Japanese researchers [3-9]. To excite a DB in graphene, a complex procedure to find proper initial conditions has been used [3]. The DB had frequency above the gapless phonon spectrum and it was proved to be unstable. The gap in the phonon spectrum of graphene can be opened by applying homogeneous elastic deformation, allowing to easily excite highly localized gap DBs [10]. Clusters of such DBs were studied in [11] and the possibility of energy exchange between DBs in clusters was demonstrated. DBs can also be excited on the edge of a stretched graphene nanoribbons of the armchair orientation [12,13]. DB frequency lies in the gap of the phonon spectrum, resulting from the application of tensile elastic strain of the nanoribbon. There exist a review on DBs in carbon and hydrocarbon nanomaterials [14] and a more general review on DBs in crystals [15].

In the two very recent papers [16,17] the authors report on the excitation of DBs in unstrained graphene in frame of the molecular dynamics simulations. In [16], using the AIREBO potential, a DB with atomic vibrations normal to the graphene sheet has been excited. Our simulations confirm this result. Even though DB frequency lies in the phonon spectrum, it is outside the spectrum of the out-of-plane phonon modes. In graphene the in-plane and out-of-plane modes interact very weakly even at relatively large amplitudes, which allows for the out-of-plane DBs to have very long lifetime. In [17], in-plane DBs in free-standing graphene have been reported based on the Tersoff interatomic potential. We will discuss in this work that the vibrational mode excited in [17] cannot be called a DB but rather it is a defect mode.

It should be pointed out that the conclusions about properties and the very existence of DBs found in molecular dynamics simulations are very sensitive to the choice of the interatomic potentials.

In this work we use the Savin potentials [18] to check what kind of DBs can be excited in the unstrained flat graphene sheet. We use the assumption that a DB can be excited by application of a bell-shape function on a short-wavelength, zone-boundary mode, whose frequency at large amplitudes leaves the phonon spectrum. This approach has been successfully used earlier [19]. We thus focus on the study of the frequency-amplitude dependence of the short-wavelength extended vibrational modes.

2. Search for DB in unstrained graphene

2.1. Phonon DOS

Phonon density of states (DOS) calculated for unstrained graphene with the use of the Savin interatomic potentials [18] is plotted in Fig. 1 for (a) in-plane modes, (b) out-of-plane modes, (c) cumulative. Highest frequency for the in-plane modes is \( \omega_{v_{max}} = 47.96 \) THz, while that for the out-of-plane modes is \( \omega_{v_{max}} = 26.94 \) THz. Symmetry of the graphene lattice precludes the appearance of a gap in the phonon DOS.

2.2. Frequencies of the in-plane short-wavelength extended modes as the functions of their amplitudes

In the work by Chechin with co-authors [20], with the use of the group-theoretical approach, four short-
wavelength vibrational modes depicted in Fig. 2 have been derived. These modes are the symmetry-dictated exact solutions to the atomic equations of motion, regardless the type of the interatomic potentials used in the simulations.

It is interesting to find the relation between amplitude and frequency for these four modes. These relations were found with the use of the Savin potential [18] and the result is presented in Fig. 3 for the modes I to IV, shown in Fig. 2 (a) to (d), respectively. Horizontal dashed line in Fig. 3 shows the upper edge of the phonon spectrum.

As it can be seen from Fig. 3, at small amplitudes, mode I has the highest frequency which bifurcates from the upper edge of the phonon spectrum and decreases with the increase in the amplitude. Mode II also demonstrates soft type nonlinearity with the mode frequency decreasing with the amplitude. Modes III and IV show hard type nonlinearity, since their frequencies increase with the amplitude. Only mode III leaves the phonon spectrum but this happens at relatively large amplitudes, greater than 0.22 Å. Since frequencies of the modes I, II, and IV lie in the phonon band, they cannot produce a DB by applying a bell-shape function upon them. We have tried to apply a bell-shape function to the mode III at amplitudes greater than 0.25 Å, but we were unable to excite a long-lived DB. At large amplitudes the instability of DB develops too fast.

2.3. Frequency of the out-of-plane short-wavelength extended mode as the function of its amplitude

Our next step is to check if the Savin potentials allow for the existence of the DBs with the out-of-plane vibrations of atoms similar to those described in [16] with the AIREBO interatomic potentials. For this we consider the out-of-plane extended vibrational mode in graphene shown in Fig. 4(a). Red and blue atoms vibrate out-of-phase in the direction normal to the graphene sheet with the amplitude A. We calculate the dependence of the mode frequency as the function of its amplitude and plot the result in Fig. 4(b), where the horizontal dashed line shows the upper edge of the phonon spectrum for the out-of-plane modes [see Fig. 1(b)]. It is clear that with the Savin potential this mode shows the soft nonlinearity type and thus, a DB with the out-of-plane atomic displacements cannot exist.

It is worth pointing out that the experimental discovery of carbon monoatomic chain [21] has inspired theoretical studies on DBs in the Fermi-Pasta-Ulam chain with atoms having three degrees of freedom [22]. The authors of the latter work have demonstrated that the well-studied longitudinal DBs become unstable in this generalized
model, but they were able to find transverse and rotational DBs and a region of the stretching force where rotational DBs are stable. Search of the transverse DBs in graphene can be regarded as a continuation of these works.

2.4. Localized mode in unstrained graphene modeled with the Tersoff interatomic potential

Fig. 4. (a) Out-of-plane vibrational mode in graphene. Red and blue atoms vibrate out-of-phase in the direction normal to the graphene sheet with the amplitude $A$. (b) Frequency as the function of amplitude for the mode shown in (a). Horizontal dashed line shows the upper edge of the phonon spectrum for the out-of-plane modes [see Fig. 1(b)].

We have revisited the results reported in [17], where the existence of DB was claimed in the free-standing graphene with the Tersoff potential. However, our simulations with the Tersoff potential have demonstrated the existence of a topological defect in graphene lattice, as shown in Fig. 5. Shown are the regular structure with translational symmetry (large black circles) and the defected structure with one valence bond, between atoms 1 and 2, longer than others (smaller gray circles). Pair of atoms 1 and 2 can vibrate out-of-phase in vertical direction with the frequency above the phonon spectrum near the defected equilibrium positions, but this mode cannot be classified as DB, because translational symmetry of the lattice is lost.

3. Conclusions

From the results presented in this work in comparison with the existing results on DBs in graphene it is clear that the effect of the interatomic potentials on the existence and properties of DBs is crucial. The reason is that the interatomic potentials are often fitted to the elastic moduli and phonon spectra of crystals (calculated from linearized equations of motion) as well as to some experimentally measurable energies, such as the sublimation energy, vacancy energy, etc. (for which not the exact profile of the potential functions but their integral characteristics are important since the change in potential energy is path independent). On the other hand, DB, being an essentially nonlinear vibrational mode, is sensitive to the exact shape of the potentials.

There exist a few studies on DBs in graphene (fully hydrogenated graphene) and strained graphene carried out in frame of the density functional theory (DFT) that takes into account electron structure of solids and does not use phenomenological interatomic potentials [23-26]. We understand that it is very important and timely to continue the DFT simulations to check the result presented in this study, particularly those for the short-wavelength modes, presented in Fig. 3 and Fig. 4. These studies are in progress and the results will be reported elsewhere.

Fig. 5. Two equilibrium configurations of graphene sheet modeled with the use of the Tersoff potential. Regular structure with translational symmetry is shown by large black circles, while smaller gray circles present a defected structure with one valence bond (between atoms 1 and 2) longer than others. Atoms neighboring to the atoms 1 and 2 have much smaller displacements from the regular lattice positions.
Acknowledgments

The authors would like to thank NOLTA2016 organizing committee members for their fruitful suggestions and comments. Financial support from the Russian Science Foundation, grant 16-12-10175, is acknowledged with thanks by S.V.D. I.P.L. thanks the grant of President of Russian Federation for Young Researchers RF MK-5283.2015.2. E.A.K. is grateful for the financial support provided by the Russian Foundation for Basic Research, grant 14-02-97029 p_povolzhie_a.

References

Discrete breathers: affecting the density shapes of heat transport

Daxing Xiong and Jun Zhang

Department of physics, Fuzhou university, Fuzhou 350108, Fujian, China
Email: phyxiongdx@fzu.edu.cn, t10077@fzu.edu.cn

Abstract—Discrete breathers (DBs), also known as intrinsic localized modes, are spatially localized nonlinear vibrational modes in anharmonic lattices, which are usually expected to affect the energy transport process. In the heat transport field of theoretical research, to predict the transport densities is currently a very fascinating topic. Inspired by these progress, we here investigate how DBs may affect the heat transport densities shapes. Through studying two peculiar one-dimensional nonlinear lattice models heat spreading processes, i.e., the Fermi-Pasta-Ulam-β chains with and without next-nearest-neighbor couplings, we will show some preliminary numerical evidences on how different types of DBs change the heat spreading densities.

1. Introduction

The studies of discrete breathers (DBs) are always hot topics of nonlinear science [1, 2, 3], among which whether DBs can contribute to heat transport in crystals is an interesting issue [4, 5, 6]. In the context of heat transport of theoretical research, recently researchers pay more attention to the densities of heat transport and its scaling behavior [7, 8, 9]. Motivated by these progress, in the present work we perform numerical simulations to investigate how DBs would affect the densities shape and their scaling law.

2. Models

We focus on the one-dimensional (1D) Fermi-Pasta-Ulam-β (FPU-β) lattices including (or not) the next-nearest-neighbor (NNN) interactions, whose Hamiltonian can be represented by

\[
H = \sum_i \left[ \frac{p_i^2}{2} + V(x_{i+1} - x_i) + \gamma V(x_{i+2} - x_i) \right], \tag{1}
\]

where \(x_i\) is the \(i\)-th particle’s displacement from its equilibrium position and \(p_i\) its momentum. The potential takes \(V(x) = \frac{1}{4}x^2 + \frac{\beta}{4}x^4\) with \(\beta\) the nonlinear parameter. Another parameter \(\gamma\) (we fix \(\gamma\) here) specifies the comparative strength of the NNN coupling to the nearest-neighbor (NN) coupling.

We note that the focused models with \(\gamma = 0\) and \(\gamma \neq 0\) will support different types of DBs, i.e., in the former model there are only extra-band DBs; while in the latter case, the intra-band DBs can be excited [4, 6].

3. Preliminary Findings

We employ the normalized spatiotemporal correlations of heat energy fluctuations [10, 11] to represent the systems heat spreading density, which is represented by

\[
\rho(x, t) = \frac{\langle \Delta Q_i(t)\Delta Q_j(0) \rangle}{\langle \Delta Q_i(0)\Delta Q_j(0) \rangle}, \tag{2}
\]

where \(\langle \cdot \rangle\) denotes the spatiotemporal average; \(\Delta Q_i(t)\) is the fluctuations of heat energy at place \(i\) and time \(t\).

Figures 1 depicts the rescaled \(\rho(x, t)\) for FPU-β chains without NNN couplings: (a) \(\beta = 0\); (b) \(\beta = 1\).

Here then is to see how these two different types of DBs would affect the heat transport densities.

Figure 1: (Color online) Rescaled \(\rho(x, t)\) for FPU-β chains without NNN couplings: (a) \(\beta = 0\); (b) \(\beta = 1\).

Figure 2: (Color online) Rescaled \(\rho(x, t)\) for FPU-β chains with NNN couplings (\(\gamma = 0.25\)): (a) \(\beta = 0\); (b) \(\beta = 0.2\).

here then is to see how these two different types of DBs would affect the heat transport densities.

- 594 -
earity ($\beta = 0$) has been introduced, there are no DBs, the density shows U shape; while if $\beta \neq 0$, the extra-band DBs emerge, the density turns to be W shape with more localization, especially on the central parts. Thus the results seem to support that the extra-band DBs will localize energy, which agrees well with the previous conjecture on the role of this type of DBs [5].

On the other hand, in [4, 6] we have suggested that compared with the extra-band DBs, the intra-band ones can be scattering with phonons, thus destroying the localization. With this picture in mind, in Fig. 2 we plot the similar results as Fig. 1 while under the systems with NNN coupling. Under this setup, the main type of DBs are the intra-band ones, then if our conjecture on intra-band DBs is still right, we should see delocalization, which has indeed been verified by Fig. 2, where the densities clearly show less localization after introducing the nonlinearity into the focused system.

Another detail is that both types of DBs appear to change the scaling [12] behaviors of the densities, i.e., when $\beta = 0$, both systems (with and without NNN couplings) show ballistic scaling behaviors; while in the case of $\beta \neq 0$, the ballistic scaling has been clearly destroyed. So one may conjecture that the DBs can not only affect the densities shapes, but also change their scaling behaviors.

4. Conclusion

To summarize, we have showed some preliminary numerical evidences on how different types of DBs affect the heat spreading densities, not only their shapes but also the scaling properties. The roles of extra-band and intra-band DBs shown in this study appear to be in good agreement with our previous conjecture, though further detailed examinations/efforts are still required on this topic.

Acknowledgments

The authors would like to thank Prof. Masayuki Kimura for his kind invitation.

References


Tunable Wave Propagation in Mechanical Metamaterials
Made of Triangulated Cylindrical Origami

Abstract for NOLTA 2016

Hiromi Yasuda and Jinkyu Yang

We study unique wave propagation in origami-based mechanical metamaterials composed of foldable origami structures, specifically Triangulated Cylindrical Origami (TCO). The TCO can be observed as a buckling pattern of a thin walled cylinder, and it can support both axial and rotational motions which can be strongly coupled with each other. To analyze wave dynamics of the TCO-based structures, we first model the TCO as a simplified 2 degree of freedom structure, by removing all facets and replacing crease lines by linear spring elements. Based on this simplified TCO unit, its static mechanical properties (e.g., force-displacement relationship and total potential energy) are examined. These mechanical properties can be controlled by altering the initial geometrical configurations of the TCO, such as height and rotational angle. In addition to the tunability of the constitutive properties, the TCO exhibits bistable behavior due to geometric nonlinearity. We verify this unique behavior experimentally by fabricating a prototype of the simplified TCO unit cell. By using the simplified TCO unit cell as a building block, we then design the TCO-based mechanical metamaterials in which multiple unit cells stacked vertically. We analyze wave propagation in this one-dimensional chain of the TCO unit cells with the focus on the coupling of axial and rotational motions. Under compressive impact, the TCO-based system exhibits unique wave dynamics such as the formation of rarefaction waves. Also, if the system consists of two distinctive TCO unit cells stacked in an alternating way, we observe formation of frequency band structures. Based on these unique characteristics, the TCO-based mechanical metamaterials have great potential for engineering applications, for example as mechanical wave filters and impact absorbers.
Localization and multimode oscillations in coupled bistable oscillators

Kuniyasu SHIMIZU†

†Dept. of Information and Communication Systems, Chiba Institute of Technology
2–17–1, Tsudanuma, Narashino, Chiba 275-0016, Japan
Email: kuniyasu.shimizu@it-chiba.ac.jp

1. Summary

A localized oscillation in coupled nonlinear media has attracted considerable attention [1]. In our previous work, we have reported several localized modes in inductively-coupled bistable oscillators by using a coupling factor as a control parameter [2]. In this study, we investigate influence of nonlinear strength on the localized modes in the coupled bistable oscillators.

A bistable oscillator is considered as a form of a van der Pol oscillator with relatively higher order nonlinearity. This oscillator consists of an inductor, capacitor, and nonlinear conductance (NC), and they are connected in parallel. The voltage–current \((v-i_{NC})\) characteristic of the NC is assumed to be given by a fifth-order polynomial:

\[
i_{NC}(v) = g_1 v - g_3 v^3 + g_5 v^5, \quad g_1, g_3, g_5 > 0.
\] (1)

The NC operates as a passive resistor when a low voltage is applied to the oscillator, and the resonance oscillation gradually decays to a non-oscillating state. On the other hand, a limit cycle oscillation exists when a high voltage is injected into the capacitor. Therefore, the oscillator has two steady states, i.e. a stable focus and a limit cycle oscillation. In this study, we investigate the bistable oscillator ring coupled by an inductor. The normalized circuit equation of the \(k\)-th bistable oscillator is written as [2]

\[
\dot{x}_k + \epsilon(1-\beta x_k^2 + x_k^4)x_k + x_k - \alpha(x_{k-1} - x_k + x_{k+1}) = 0,
\] (2)

where \(N\) is the number of hard oscillators. In the following results, the phenomena observed in six inductor-coupled hard-oscillator rings \((N = 6)\) are investigated for simplicity.

Figure 1 shows the time series for \(\epsilon = 0.39, \alpha = 0.1, \) and \(\beta = 3.2\). Initial conditions are \(x_1 = 2.0, x_2 = 0.5, x_3 = 0.1, x_4 = 0.2, x_5 = 0.9, x_6 = -2.2, y_1 = 1.8, y_2 = -0.3, y_3 = y_4 = 0.3, y_5 = -2.3, \) and \(y_6 = 0.4\).

Figure 2: Propagating wave in the six inductor-coupled hard oscillators for \(\epsilon = 0.38, \alpha = 0.1, \) and \(\beta = 3.2\). The initial conditions are the same as those in Fig. 1.

References


Extensions of a Theorem on Algebraic Connectivity Maximizing Graphs

Ryoya Ishii and Norikazu Takahashi
Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama, 700–8530 Japan
Email: ishii@momo.cs.okayama-u.ac.jp, takahashi@cs.okayama-u.ac.jp

Abstract—The second smallest eigenvalue of the Laplacian matrix of a graph, also known as the algebraic connectivity, is an important measure that represents how strongly the graph is connected. The algebraic connectivity also characterizes the performance of some dynamic processes on networks such as consensus in multiagent networks and synchronization of coupled oscillators. In this paper, we study the problem of finding graphs that maximize the algebraic connectivity among all graphs with the same number of vertices and edges, and extend a known result about complete bipartite graphs to complete multipartite graphs.

1. Introduction

Algebraic connectivity [1] of a graph, which is defined as the second smallest eigenvalue of the Laplacian matrix, is an important measure that represents how strongly the graph is connected. Not only it has been intensively studied in mathematics [2, 3, 4], but also it has attracted a great deal of attention from researchers in engineering. For example, the convergence rate of a well-known consensus algorithm for multiagent networks is determined by the algebraic connectivity among all graphs with the same degree matrix.

2. Algebraic Connectivity Maximizing Graphs

2.1. Notations and Definitions

Throughout this paper, by a graph we mean a simple undirected graph. Let \( G = (V(G), E(G)) \) be a graph with the vertex set \( V(G) = \{1, 2, \ldots, n\} \) and the edge set \( E(G) \).

The Laplacian matrix of \( G \) is defined by \( L(G) = D(G) - A(G) \) [1] where \( D(G) \) is the degree matrix and \( A(G) = (a_{ij}(G)) \in \{0, 1\}^{n \times n} \) is the adjacency matrix. Because \( L(G) \) is positive semi-definite, its eigenvalues, which are denoted by \( \lambda_1(G), \lambda_2(G), \ldots, \lambda_n(G) \), are nonnegative real numbers. In the remainder of this paper, we assume without loss of generality that \( 0 \leq \lambda_1(G) \leq \lambda_2(G) \leq \cdots \leq \lambda_n(G) \).

Recently, Ogiwara et al. [8] studied the problem of finding graphs with a given number of vertices and edges that maximize the algebraic connectivity. This problem is important from various perspectives such as the fast convergence of the consensus algorithm, the robustness of networks against failures and attacks, and so on. They proved that some well-known classes of graphs such as star graphs, cycle graphs and complete bipartite graphs are algebraic connectivity maximizers under certain conditions. This problem was also considered by Kolokolnikov [9]. He presented a conjecture that the complete bipartite graph \( K_{2,n-2} \) maximizes the algebraic connectivity among all graphs with \( n \) vertices and \( 2(n-2) \) edges. He also showed by exhaustive search that this conjecture holds true for all \( n \) less than or equal to 13. Fujihara and Takahashi [10] studied a slightly different problem and proved that any complete multipartite graph maximizes the algebraic connectivity among all graphs with the same degree matrix.

In this paper, we prove that if a complete multipartite graph satisfies a certain condition then it maximizes the algebraic connectivity among all graphs with the same number of vertices and edges. This is an extension of a theorem given by Ogiwara et al. [8], which states that any complete bipartite graph \( K_{n_1,n_2} \) with \( n_1 \leq n_2 \) maximizes the algebraic connectivity among all graphs with \( n_1+n_2 \) vertices and \( n_1n_2 \) edges. We further generalize this result to graphs obtained from complete multipartite graphs by adding some edges.

Definition 1 Let \( G_{n,m} \) be the set of all graphs with \( n \) vertices and \( m \) edges. If a graph \( G \in G_{n,m} \) satisfies the condition that

\[
\forall G' \in G_{n,m}, \quad \lambda_2(G) \geq \lambda_2(G')
\]

then \( G \) is called an algebraic connectivity maximizing graph in \( G_{n,m} \).

2.2. Known Results

If the vertex set \( V(G) = \{1, 2, \ldots, n\} \) of a graph \( G \) is partitioned into \( k \geq 2 \) subsets \( V_1, V_2, \ldots, V_k \) in such a way that vertices \( i \in V_a \) and \( j \in V_b \) are adjacent to each other if and only if \( a \neq b \), then \( G \) is called a complete \( k \)-partite graph and denoted by \( K_{n_1,n_2,\ldots,n_k} \) where \( n_l = |V_l| \) for \( l = 1, 2, \ldots, k \). An example of such a graph is shown in Fig. 1.
In the following discussions, we assume without loss of generality that
\[ 1 \leq n_1 \leq n_2 \leq \cdots \leq n_k . \tag{1} \]

The complete \( n \)-partite graph \( K_{1,1,\ldots,1} \) is called the complete graph and denoted by \( K_n \) in this paper.

First, we present some fundamental results about the eigenvalues of the Laplacian matrix and their multiplicities.

**Lemma 1** ([2]) The eigenvalues of \( L(K_n) \) are 0, with multiplicity 1, and \( n \), with multiplicity \( n-1 \).

**Theorem 1** ([2]) If \( \lambda \) is an eigenvalue of \( L(G) \) then \( 0 \leq \lambda \leq n \). The multiplicity of \( 0 \) is the number of connected components of \( G \). The multiplicity of \( n \) equals one less than the number of connected components of \( G^c \), the complement of \( G \).

**Theorem 2** The eigenvalues of \( L(K_{n_1,n_2,\ldots,n_k}) \) are 0, \( n - n_k, n - n_{k-1}, \ldots, n - n_1 \) and \( n \), with multiplicity 1, \( n_k - 1, n_{k-1} - 1, \ldots, n_1 - 1 \) and \( k - 1 \), respectively.

**Proof:** Because \( K_{n_1,n_2,\ldots,n_k} \) is connected, it follows from Theorem 1 that the smallest eigenvalue 0 of \( L(K_{n_1,n_2,\ldots,n_k}) \) has multiplicity 1. Also, because the complement \( K_{n_k,n_{k-1},\ldots,n_1} \) of \( K_{n_1,n_2,\ldots,n_k} \) has connected components which are isomorphic to \( K_{n_1}, K_{n_2}, \ldots, K_{n_k} \), it follows from Theorem 1 that \( L(K_{n_1,n_2,\ldots,n_k}) \) has the largest eigenvalue \( n \) with multiplicity \( k - 1 \). Moreover, we see from Lemma 1 that the eigenvalues of \( L(K_{n_1,n_2,\ldots,n_k}) \) are 0, \( n_1, n_2, \ldots, n_k \) with multiplicity \( n_k - 1, n_{k-1} - 1, \ldots, n_1 - 1 \), respectively. Hence, in order to complete the proof, we only have to show that if \( L(K_{n_1,n_2,\ldots,n_k}) \) has an eigenvalue \( \lambda \) other than 0 and \( n \) with multiplicity \( \mu \) then \( L(K_{n_1,n_2,\ldots,n_k}) \) has an eigenvalue \( n - \lambda \) with the same multiplicity.

Note that \( K_{n_1,n_2,\ldots,n_k} \) and \( K_{n_k,n_{k-1},\ldots,n_1} \) satisfy
\[ L(K_{n_1,n_2,\ldots,n_k}) = L(K_n) - L(K_{n_k,n_{k-1},\ldots,n_1}). \tag{2} \]

Let \( \nu \) be an eigenvector of \( L(K_{n_k,n_{k-1},\ldots,n_1}) \) associated with \( \lambda \). Then \( \nu \) is orthogonal to 1. Multiplying both sides of (2) by \( \nu \) from right, we have
\[ (L(K_n) - L(K_{n_k,n_{k-1},\ldots,n_1}))\nu = (nI - 11^T)\nu - \lambda \nu = \nu \nu - \lambda \nu = (n - \lambda)\nu \tag{3} \]

where \( I \) is the identity matrix. From (2) and (3) we have
\[ L(K_{n_1,n_2,\ldots,n_k})\nu = (n - \lambda)\nu \]

which means that \( n - \lambda \) is an eigenvalue of \( L(K_{n_1,n_2,\ldots,n_k}) \) and \( \nu \) is an eigenvector associated with \( n - \lambda \). In addition, it is easy to see that the dimension of the eigenspace of \( L(K_{n_1,n_2,\ldots,n_k}) \) associated with \( n - \lambda \) is equal to that of \( L(K_{n_k,n_{k-1},\ldots,n_1}) \) associated with \( n - \lambda \).

Theorem 2 may not be new. However, it is difficult to find this result in the existing literature. We therefore have provided a proof of it.

Next, we present two results given by Ogiwara *et al.* [8] about the sufficient condition for a complete bipartite graph to be an algebraic connectivity maximizing graph.

**Theorem 3** ([8]) If \( k = 2 \) and two positive integers \( n_1 \) and \( n_2 \) satisfy \( n_1 + n_2 \geq 3 \) and
\[ n_1 = \frac{2n_1^2}{n_1 + n_2} < 1 \]
as well as (1) then the complete bipartite graph \( K_{n_1,n_2} \) is an algebraic connectivity maximizing graph in \( G_{n_1 + n_2,n_1n_2} \).

**Corollary 1** ([8]) If \( k = 2 \) and two positive integers \( n_1 \) and \( n_2 \) satisfy \( n_1 + n_2 \geq 3 \) and
\[ \left[ \frac{n_1 + n_2 - 1}{2} \right] \leq n_1 \leq \left[ \frac{n_1 + n_2}{2} \right] \]
as well as (1) then the complete bipartite graph \( K_{n_1,n_2} \) is an algebraic connectivity maximizing graph in \( G_{n_1 + n_2,n_1n_2} \).

We also provide two well-known results about the algebraic connectivity, that will be needed in later discussions.

**Theorem 4** ([11]) If \( G \) is not a complete graph then \( \lambda_2(G) \leq \delta(G) \) where \( \delta(G) = \min_{1 \leq i \leq |V(G)|} \{ \delta_i(G) \} \).

**Theorem 5** ([3]) Let \( G' \in G_{n,m+1} \) be a graph obtained by adding an edge to \( G \in G_{n,m} \). Then we have
\[ \lambda_1(G) \leq \lambda_1(G') \leq \lambda_2(G) \leq \lambda_2(G') \leq \cdots \leq \lambda_q(G) \leq \lambda_q(G') \]

3. Exhaustive Search of Algebraic Connectivity Maximizing Graphs

In order to see what kind of complete multipartite graphs can be algebraic connectivity maximizing graphs, we developed an exhaustive search algorithm based on the graph enumeration algorithm proposed by Sato and Nakano [11] and applied it to \( G_{n,m} \) for various values of \( (n, m) \).

We first applied the algorithm to \( G_{n,m} \) with \( n \leq 10 \) such that it contains a complete bipartite graph. For \( n = 10 \), for example, the algorithm was applied to \( G_{10,9}, G_{10,16}, G_{10,21}, G_{10,24}, \) and \( G_{10,25} \). As a result, it was found that any bipartite graph \( K_{n_1,n_2} \) with \( n_1 + n_2 \leq 10 \) is an algebraic connectivity maximizing graph in \( G_{n_1 + n_2,n_1n_2} \).
Next we applied the algorithm to $G_{8,17}$ such that a complete tripartite graph is contained in it. First let us consider $G_{7,14}$ which contains $K_{1,2,4}$. An algebraic connectivity maximizing graph found by the algorithm is shown in Fig. 2. Note that it is not a complete multipartite graph. Moreover, all other graphs found by the algorithm were isomorphic to the graph in Fig. 2. This means that $K_{1,2,4}$ is not an algebraic connectivity maximizing graph. On the other hand, for $G_{8,24}$ and $G_{9,27}$, the algorithm found $K_{2,2,2,2}$ and $K_{3,3,3,3}$, respectively, as algebraic connectivity maximizing graphs. From these results and Corollary 1, it is conjectured that the complete $k$-partite graph $K_{n_1,n_2,...,n_k}$ with $n_1 = n_2 = \cdots = n_k$ is an algebraic connectivity maximizing graph. It is proved in the next section that the conjecture is in fact true.

We also applied the algorithm to $G_{n,m}$ which does not necessarily contain a complete multipartite graph. An algebraic connectivity maximizing graph found for $G_{8,17}$ is shown in Fig. 3 and that for $G_{9,28}$ is shown in Fig. 4. It is easily seen that each of them is obtained from a complete multipartite graph $K_{n_1,n_2,...,n_k}$ with $n_1 = n_2 = \cdots = n_k$ by adding one edge. It is proved in the next section that these graphs are algebraic connectivity maximizing graphs.

4. Theoretical Analysis

We give two theorems that can be considered as extensions of Theorem 3 and Corollary 1. Before doing so, we present two lemmas.

**Lemma 2** If $k \geq 2$ and $k$ positive integers $n_1, n_2, \ldots, n_k$ satisfy

$$n_k \geq 2 \quad \text{and} \quad n_k \sum_{i=1}^{k-1} n_i \leq \sum_{i=1}^{k-1} n_i^2$$

as well as (1) then the complete $k$-partite graph $K_{n_1,n_2,...,n_k}$ is an algebraic connectivity maximizing graph found by the exhaustive search algorithm for $G_{n,m}$ where $n = \sum_{i=1}^{k} n_i$ and $m = \sum_{i=1}^{k} n_i(n - n_i)/2$.

**Proof:** By Theorem 2, the algebraic connectivity of the complete $k$-partite graph $K_{n_1,n_2,...,n_k}$ is equal to $n - n_k$. We therefore prove under the assumption (4) that $\lambda_2(G) \leq n - n_k$ for all $G \in G_{n,m}$ where $n = \sum_{i=1}^{k} n_i$ and $m = \sum_{i=1}^{k} n_i(n - n_i)/2$. Furthermore, by Theorem 4, it suffices for us to show under the assumption (4) that $\delta(G) \leq n - n_k$ for all $G \in G_{n,m}$ (note that $G$ is not a complete graph because of the assumption $n_k \geq 2$). The sum of the degrees of all vertices of $G$ is given by

$$\sum_{j=1}^{n} d_j(G) = 2m = \sum_{i=1}^{k} n_i(n - n_i) = n^2 - \sum_{j=1}^{n} n_j^2.$$

Here it follows from assumption (4) that

$$- \sum_{k} n_i^2 = - \sum_{k} n_i^2 - n_k^2 \leq -n_k \sum_{i=1}^{k} n_i - n_k^2 = -nn_k$$

from which we have

$$\sum_{i=1}^{n} d_i(G) \leq n^2 - nn_k = n(n - n_k).$$

Therefore, we finally have

$$\delta(G) \leq \frac{1}{n} \sum_{i=1}^{n} d_i(G) = n - n_k$$

which completes the proof. \qed

**Lemma 3** Let $k$ be any integer greater than or equal to 2. Positive integers $n_1, n_2, \ldots, n_k$ satisfy (1) and (4) if and only if $n_1 = n_2 = \cdots = n_k \geq 2$.

**Proof:** It follows from (1) that $n_k n_1 \geq n_i^2$ for $l = 1, 2, \ldots, k - 1$. Hence (4) holds if and only if $2n_1 \leq n_k n_1 = n_i^2$ for $l = 1, 2, \ldots, k - 1$, that is, $n_1 = n_2 = \cdots = n_k \geq 2$. \qed

From Lemmas 2 and 3, we immediately obtain the following theorem.

**Theorem 6** If $k \geq 2$, the positive integers $n_1, n_2, \ldots, n_k$ are equal to each other, and $n_1 \geq 2$ then the complete $k$-partite graph $K_{n_1,n_2,...,n_k}$ is an algebraic connectivity maximizing graph in $G_{n,m}$ where $n = kn_1$ and $m = kn_1(n - n_1)/2$. 

- 600 -
This theorem can be further extended as follows.

**Theorem 7** If \( k \geq 2 \), the positive integers \( n_1, n_2, \ldots, n_k \) are equal to each other, \( n_1 \geq 2 \), and \( p \) is a positive integer less than \( kn_1/2 \) then any graph obtained from the complete \( k \)-partite graph \( K_{n_1,n_2,...,n_k} \) by adding \( p \) edges has the same algebraic connectivity as \( K_{n_1,n_2,...,n_k} \) and is an algebraic connectivity maximizing graph in \( G_{n+m,p} \) where \( n = kn_1 \) and \( m = n(n-n_1)/2 \).

**Proof:** Suppose that \( k, n_1, n_2, \ldots, n_k \) and \( p \) satisfy the assumptions of the statement. We first show that

\[
p \leq k(n_1 - 1) - 1.
\]

If \( kn_1 \) is even then we have

\[
k(n_1 - 1) - 1 - p \geq k(n_1 - 1) - 1 - \left( \frac{kn_1}{2} - 1 \right) = \frac{k}{2}(n_1 - 2)
\]

which is nonnegative. If \( kn_1 \) is odd then we have

\[
k(n_1 - 1) - 1 - p \geq k(n_1 - 1) - 1 - \frac{kn_1 - 1}{2} = \frac{k(n_1 - 2) - 1}{2}
\]

which is positive because \( k \geq 3 \) and \( n_1 \geq 3 \). Therefore, \( p \) always satisfies (5). Let \( G' \) be any graph obtained from the complete \( k \)-partite graph \( K_{n_1,n_2,...,n_k} \) by adding \( p \) edges. The eigenvalues of \( L(K_{n_1,n_2,...,n_k}) \) are 0 with multiplicity 1, \( n-n_1 \) with multiplicity \( k(n_1 - 1) \), and \( n \) with multiplicity \( k-1 \) due to Theorem 2. By this fact, together with Theorem 5 and (5), we have \( \lambda_2(G') = n - n_1 = \lambda_2(K_{n_1,n_2,...,n_k}) \). In order to prove the second part, it suffices to show that \( \lambda_2(G) \leq n-n_1 \) for any \( G \in G_{n+m,p} \). By Theorem 4, we have

\[
\lambda_2(G) \leq \delta(G) \leq \frac{1}{n} \sum_{i=1}^{n} d_i(G) = \frac{2(m+p)}{n}
\]

where

\[
\frac{2(m+p)}{n} = \frac{n(n-n_1)+2p}{n} = n-n_1 + \frac{2p}{n}
\]

and \( 2p/n \) is less than 1 from the assumption. Therefore, the inequality (6) implies that \( \lambda_2(G) \leq n-n_1 \). \( \square \)

5. Conclusion

In this paper, we first proved that any complete multipartite graph \( K_{n_1,n_2,...,n_k} \) with \( n_1 = n_2 = \cdots = n_k \) is an algebraic connectivity maximizing graph. We then extended this result to graphs obtained from such complete multipartite graphs by adding some edges. However, we have to say that these results are rather severe, because Theorems 6 or 7 apply to the set of graphs with \( n \) vertices and \( m \) edges for relatively few choices of \( n \) and \( m \). A future problem is to obtain milder sufficient conditions.

**Acknowledgement**

This work was partially supported by JSPS KAKENHI Grant Number JP15K00035.

**References**


Theory for Dynamical Robustness of Complex Networks against Targeted Attacks

Tianyu Yuan†, Kazuyuki Aihara†,‡, Gouhei Tanaka†,‡

†Department of Electrical Engineering and Information Systems, the University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
‡Institute of Industrial Science, the University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan
Email: yty@sat.t.u-tokyo.ac.jp, aihara@sat.t.u-tokyo.ac.jp, gouhei@sat.t.u-tokyo.ac.jp

Abstract— Many real-world networks do not have simple static structures but dynamics on them, such as many social, biological and computer networks. Such networks are heterogeneously connected. It is meaningful to investigate the dynamical robustness of complex networks against random failures and targeted attacks. So far the theory for analyzing the dynamical robustness against random failures has been developed, but that against targeted attacks is missing. This paper derives an analytical expression for the critical value which measures the dynamical robustness against targeted attacks for a general kind of coupled dynamical networks.

1. Introduction

Real-world networks are often heterogeneous, which have a wide range of degree distribution. One famous case is the scale-free network with power law degree distribution [1]. Such networks are known to be robust against random failures but very fragile to targeted attacks at hubs [2], which means preferential removal of high-degree nodes. This is the result of the analysis in the framework of structural robustness where only connectivity of the networks governs the robustness. The theoretical treatment has been based on the percolation theory [3, 4, 14]. However, it is also necessary to study dynamical robustness of complex networks, which is especially important in biological phenomena because the function of biological networks is usually considered to depend on both structure and dynamics [5].

In this paper we deal with dynamical robustness of diffusively coupled oscillator networks with complex topology. The dynamical robustness represents how the dynamical behavior in the whole network is tolerant against local failures. In the dynamical robustness framework, the local failures correspond to inactivation of the oscillator nodes. The fraction of the inactivated nodes is denoted by \( p \). The order parameter is introduced to measure the level of the oscillatory behavior in the whole network and the critical inactivation ratio \( p_c \), at which a loss of oscillatory dynamics (aging transition [6]) occurs, is used to measure the network robustness. This framework for study of oscillator networks has been first introduced for globally coupled networks in Ref. [6] and subsequently extended to complex networks [7-11]. These studies have employed the Stuart-Landau oscillators which generate oscillations via the Hopf bifurcation. The most striking result of the dynamical robustness analysis is that the oscillator network can be highly fragile to the attack targeted at low-degree oscillator nodes, instead of high-degree hub nodes [7]. The critical value \( p_c \) has been analytically obtained for random failures [7], but not for targeted attacks. The main work of this paper is to give the theoretical \( p_c \) for targeted attacks, which works for a general diffusively coupled oscillators including the Stuart-Landau model. It is shown analytically that the property of dynamical robustness is different from the structural robustness, indicating that the effects of high-degree attack, low-degree attack, and random failures depend on the range of parameters.

2. Methods

We consider a general kind of diffusively coupled oscillators as below:

\[
\dot{z}_j = F(z_j) + h_j \sum_{k=1}^{N} A_{jk} (z_k - z_j),
\]

\[
\text{for } j = 1, ..., N, \tag{1}
\]

where \( F(z_j) \) is the function representing the dynamics of individual nodes, exhibiting a bifurcation; \( N \) is the number of oscillators; \( z_j \) is the complex value representing the state of the \( j \)th oscillator; \( A_{jk} \) is the adjacency matrix which takes any real value and does not have to be symmetric; \( h_j \) is a tuning term. If \( h_j \) depends on \( A_{jk} = \sum_{k=1}^{N} A_{jk} \), then it becomes a weighted network, which is studied in Ref. [8]. Although \( z_j \) is supposed to be complex, we assume based on numerical observation that all nodes tend to have the identical phases, so \( z_j \) can be viewed as real.

A widely used special case is described as follows [6]:

\[
\dot{z}_j = \left( a_j + i \Omega - |z_j|^2 \right) z_j + \frac{K}{N} \sum_{k=1}^{N} A_{jk} (z_k - z_j),
\]

\[
\text{for } j = 1, ..., N, \tag{2}
\]

where \( \Omega \) is the natural frequency; \( K \) is the coupling strength. Here \( A_{jk} \) takes the value of 0 or 1. Later we will consider model (2) as well as a weighted coupling case described as follows [8]:
\[ \dot{z}_j = (\alpha_j + i\Omega - |z_j|^2)z_j + \frac{K(k)}{NK_j} \sum_{k=1}^{N} A_{jk}(z_k - z_j), \]

where \( k_j = \sum_{k=1}^{N} A_{jk} \) is the degree of the \( j \)th node and \( \langle k \rangle \) is the average degree. An isolated oscillator \((K=0)\) is called the Stuart-Landau oscillator, which represents the normal form of Hopf bifurcation which occurs at \( \alpha_j = 0 \). An oscillator is called active when \( \alpha_j = a > 0 \) and inactive when \( \alpha_j = -b < 0 \). An inactive oscillator tends to approach the origin with \( |z_j| = 0 \) after transient damped oscillations while an active oscillator tends to the limit cycle with amplitude \( |z_j| = \sqrt{a} \). Such notations can also be used for the general case (1): an active oscillator with \( F(z_j) = F_A(z_j) \) and an inactive one with \( F(z_j) = F_I(z_j) \). Therefore random failures and targeted attacks can be defined as inactivation of active oscillators.

The order parameter \( |Z| \) is defined to measure the level of oscillation of the network: \( |Z| = \frac{1}{p} \sum_{j=1}^{N} z_j \). As the ratio \( p \) of inactive oscillators increases from 0 to 1, the order parameter \( |Z| \) decreases and will reach 0 at some critical ratio \( p_c \). We can use \( p_c \) as a measure of dynamical robustness. To calculate the critical ratio \( p_c \), we assume the heterogeneous (degree-weighted) mean field approximation [7, 12]. Moreover we remove the phase rotation of all nodes and assume that every \( z_j \) is real. Let \( A_j = \sum_{k=1}^{N} A_{jk} \) and then

\[ H_A(t) = \sum_{j \in S_A} A_j z_j(t), \quad H_I(t) = \sum_{j \in S_I} A_j z_j(t), \]

where \( S_A \) and \( S_I \) are the sets of active and inactive nodes, respectively. Then the mean-field approximation gives

\[ \sum_{j=1}^{N} A_{jk} z_k \cong A_j \frac{\sum_{k \in S_A} A_{jk} z_k(t)}{\sum_{k \in S_A} A_{jk}} = A_j \frac{H_A(t) + H_I(t)}{\sum_{k \in S_A} A_{jk}} - z_j = 0. \]

This equation assumes that every link connects to any node with the same probability, i.e. the network is uncorrelated. By substituting Eq. (5) into Eq. (1) and considering the condition for an equilibrium, we obtain

\[ \dot{z}_j = F(z_j, H_A, H_I) = \frac{F_A(z_j)}{H_A(t) + H_I(t)} - z_j = 0. \]

Then, it follows

\[ \frac{\partial z_j}{\partial H_A} = \frac{\partial z_j}{\partial H_I} = \frac{\partial z_j}{\partial H_K} = \frac{\partial z_j}{\partial H_{\theta}} = 0, \]

where

\[ F(z_j) = \begin{cases} F_A(z_j) & \text{for active nodes} \\ F_I(z_j) & \text{for inactive nodes} \end{cases} \]

The derivative of Eq. (4) to \( H_A(t), H_I(t) \) becomes

\[ J = \begin{bmatrix} \frac{\partial}{\partial H_A} (\sum_{j \in S_A} A_j z_j(t)) - 1 & \frac{\partial}{\partial H_I} (\sum_{j \in S_A} A_j z_j(t)) \\ \frac{\partial}{\partial H_I} (\sum_{j \in S_A} A_j z_j(t)) & \frac{\partial}{\partial H_I} (\sum_{j \in S_A} A_j z_j(t)) - 1 \end{bmatrix}. \]

From the inverse function theorem, (4) has a derivable solution if and only if \( J \) is nonsingular. It is easy to see that

\[ \sum_{j=1}^{N} A_j = \sum_{j \in S_A} A_j + \sum_{j \in S_I} A_j = \sum_{j \in S_A} A_j + \sum_{j \in S_I} A_j. \]

For random failures, Eq. (8) becomes

\[ \sum_{j=1}^{N} A_j = \sum_{j \in S_A} A_j + A_j^2 \sum_{j \in S_I} A_j \]

and thus

\[ \sum_{j=1}^{N} A_j = \sum_{j \in S_A} A_j + A_j^2 \sum_{j \in S_I} A_j. \]

For targeted attacks, we sort the indices of oscillators by the preferential order of attacks, and thus the oscillators \( j = 1, ..., N_{pc} \) are inactive and \( j = N_{pc} + 1, ..., N \) are active. Since the right-hand side of Eq. (8) is supposed to be a monotone function of \( p_c \), there has to be a solution \( p_c \) satisfying the following equation

\[ \sum_{j=1}^{N} A_j = \sum_{j \in S_A} A_j + A_j^2 \sum_{j \in S_I} A_j. \]

The solution \( p_c \) is numerically found in practice. For the coupled Stuart-Landau oscillator model in Eq. (2), the expressions become as below:

For random failures:

\[ p_c = \frac{\sum_{j=1}^{N} k_j}{\sum_{j=1}^{N} k_j + \sum_{j \neq k} k_j}. \]

This formula is consistent with the previous result [7].

For targeted attacks:

\[ \sum_{j=1}^{N} k_j = \sum_{j \in S_A} k_j + \sum_{j \in S_I} k_j \]

For the weighted coupling model in Eq. (3), it turns out to be very simple:

For random failures:

\[ p_c = \frac{1 + \frac{BN}{K}}{1 + \frac{BN}{K}}. \]

For targeted attacks:

\[ N = \frac{1}{(k+BN)/K} \sum_{j=1}^{N_{pc}} k_j + \frac{1}{(k+BN)/K} \sum_{j \neq k} k_j. \]

Obviously in the case of weighted coupling networks the low-degree attack always gives the higher \( p_c \) and the high-degree attack gives the lower \( p_c \).

3. Results

In this section we show that the theoretically derived \( p_c \) is validated in numerical simulations for the Stuart-Landau model and the weighted coupling Stuart-Landau model. Numerical computation is run by the Matlab’s built-in function ode45, a four-five order Runge-Kutta method. All results are sampled after running 1000 time steps. For targeted attacks we consider the special cases of attacks.
targeted at the high-degree nodes (high-degree attack) and at the low-degree nodes (low-degree attack). Therefore the sorting of attack preference is just sorting by degrees.

Figure 1(a) shows the time courses of the oscillatory dynamics after a sufficiently long time, for the whole oscillators, the active group, and the inactive group. Each curve represents the real part of the state variables averaged over each group. The mean value of all oscillators gets stable at time 500. Figure 1(b) shows the procedure of the aging transition for a randomly connected network. The behavior of the curve near $p_c$ is studied in Ref. [6]. In this case, the network is more vulnerable to the low-degree attack than to the high-degree attack.

Figure 2 shows the critical ratio $p_c$ vs coupling strength $K$. The numerical results match well with the theoretically computed $p_c$. In Fig. 2(a), the low-degree attack leads to the smaller $p_c$ compared with the random failure and the high-degree attack for the whole range of $K$. However, this property does not hold when the parameter values of $a$ and $b$ are changed. From Eq. (13), it is easy to see that when $a$ is very small, the high-degree attacks will finally have lower $p_c$, as shown in Fig. 2(b) with $a=0.1$ and $b=1$.

Figure 3 shows the weighted coupling case. As predicted by Eq. (15), the high-degree attack gives the lowest $p_c$ while the low-degree attack gives the highest, which matches the result in [8].

Figure 1. (a)(b) Mean values of oscillators vs time for different $p$. (a) is the case of random failure at $p=0.6$, (b) is the damped case of random failure at $p=0.9$. (c) The order parameter $|Z|$ vs inactivation ratio $p$. The parameters are set at $N=100$, connection density $d=0.3$, $a=b=1$, $\Omega=0.1$, and $K=5$.

Figure 2. The critical ratio $p_c$ vs the coupling strength $K$. The number of oscillators is $N=1000$; the scale-free network is built by Barabási–Albert method with preferential attachment of 40 links each step. $\Omega=0.1$. (a) $a=b=1$; (b) $a=0.1$, $b=1$. 

- 604 -
Figure 3. The critical ratio $p_c$ vs the coupling strength $K$ for the weighted network. The number of oscillators is $N=1000$; the scale-free network is built by Barabási–Albert method with preferential attachment of 40 links each step. $\Omega=0.1$. (a) $a=b=1$; (b) $a=0.1$, $b=1$.

4. Conclusions

In this study, we have derived the theoretical formula for the critical value of $p_c$ for targeted attacks in a general network model of diffusively coupled bifurcating systems. The theoretical results are in good agreement with the numerical ones in the scale-free networks of coupled Stuart-Landau oscillators and the weighted coupling model. It has been demonstrated that the attack method which is most dangerous for network robustness can change depending on the dynamics of the individual components of the network.

The theoretically derived critical values in Eqs. (9)-(10) are not confined to these cases and should work for other types of targeted attacks. However, the computation of the critical ratio requires the mean-field approximation, thus the theoretical $p_c$ may be problematic for networks with high-order connection, such as correlated networks [13]. Since the theoretical $p_c$ only depends on the degree distribution and attack preference, it has no high-order information of the network.

Future works will consider other types of dynamical components as well as different preferential attacks. Moreover, it is worth extending the study to correlated networks and pulse-coupled networks.

References

A Network Analysis of World’s Metro Systems

Xingtang Wu†, Chi K. Tse‡, Hairong Dong†, Ivan W. H. Ho‡ and Francis C. M. Lau‡

†School of Electronic and Information Engineering, Beijing Jiaotong University, China
‡Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong, China

Email: 14111051@bjtu.edu.cn, cktse@ieee.org

Abstract—Metro systems carry a large volume of commuters around in major cities and their efficiency depends on multiple factors including the connectivity of subway lines, train schedules, passenger distributions, management quality, etc. In this paper, we study the network properties of five metro systems (Beijing, Hong Kong, London, Paris and Tokyo) and compare their relative network efficiency in terms of two proposed metrics, namely, average station density and station load. It is shown that among the five systems, the Tokyo system has the shortest characteristic path length (shortest average travel distance between two stations), as well as highest efficiency in carrying passengers around the city. Furthermore, the London metro has a better tolerance to faults in a local scale, and the Paris system outperforms others in terms of level of convenience to commuters due to its high station density and low load.

1. Introduction

Rapid transit systems, often called metro or subway systems, are transportation systems carrying the largest volume of commuters in major cities, and their reliability, efficiency, safety, levels of comfort, convenience and accessibility are often perceived by travellers and local commuters as indicators of the quality of public transportation of the cities [1]. Major cities, due to increasing traffic demands and ever-extending city coverage, are continuously expanding their metro networks, resulting in complex subway systems that possess high station densities and intricate inter-station couplings [2]. Design and scheduling of metro systems to optimize performance has become important considerations in the development of public transportation systems. Moreover, the study of networks, under the notion of complex networks, has recently become popular due to the intriguing discovery of a number of universal properties in various physical and man-made networks [3, 4] and the promising applications that have been developed in various practical fields such as communications, power systems, finance, disease control, etc. [5]–[10]. Results from complex networks research are highly relevant to the study of transportation, especially in the provision of appropriate analytical tools for characterizing the structure of metro systems which are practical forms of networks and for understanding the operations of a complex system such as metro systems [11, 12]. Furthermore, the huge investment in this transportation infrastructure and the impact to the public certainly justify a more thorough investigation of the factors affecting performance, thus allowing a more informed planning and design for future development.

The cross-disciplinary study of subway systems from a perspective of complex networks is still relatively rare. The earliest work reported by Latora and Marchiori [13] showed that the Boston subway network exhibited the small-world property and introduced the concept of network efficiency to give useful insights on the general characteristics of real transportation networks. In the work of Derrible and Kennedy [14], most metros were found to exhibit scale-free and small-world structure. Also, Angeloudis and Fisk [2] studied 20 subway networks using a ‘toy’ model and showed that these networks, with high connectivity and low maximum vertex degrees, provide robustness to random attacks. In the work of Lee et al. [15], the statistical properties of the Metropolitan Seoul subway network were analyzed, taking the passenger flow as the weight of the edge and arriving at a power-law weight distribution. Furthermore, Yang et al. [5] combined node degree and betweenness to assess the node importance, and showed that a scale-free transit network exhibited a relatively high fault tolerance to random failure but a relatively low degree of connection reliability against malicious attack.

In this paper, five subway networks are studied via analyzing some network parameters such as degree distribution and network efficiency, the aim being to identify the factors affecting performance.

2. Topological properties of subway network

A complex network with $N$ nodes can be represented as a graph $G = (N, l)$, where $N = [n_1, n_2, ..., n_N]$ denotes the set of nodes, and $l = [l_1, l_2, ..., l_L]$ denotes the set of links. A graph $G$ can be fully described by an adjacency matrix $A$, which is an $N \times N$ matrix whose entry $a_{ij}$ ($i, j = 1, ..., N$) equals to 1 if there exists a link between nodes $i$ and $j$, and zero otherwise. In this paper, a node is a subway station, if two stations are directly connected by a track, then they are connected by a link.

2.1. Characteristic path length

Shortest path length, denoted as $d_{ij}$, is the shortest length from nodes $i$ to $j$, which plays an important role in
transportation and communication networks. Suppose one needs to commute from one station to another by subway: the shortest path provides an optimal pathway in the sense that one would achieve a fast transfer, saving time and resources. A measure of the typical separation between two nodes in a complex network is given by the characteristic path length, also known as average path length, which is defined as the mean shortest path lengths over all pairs of nodes [3]:

$$L = \frac{1}{N(N-1)} \sum_{i<j} d_{ij}$$

(1)

This parameter directly indicates the global connectivity of a network. A smaller value of $L$ represents smaller topological distance between any two nodes and better connectivity of the whole network.

2.2. Clustering coefficient

Clustering coefficient $C$, also known as transitivity, is a typical property of acquaintance networks, where two individuals with a common friend may know each other. One definition of $C$, introduced by Watts and Strogatz [3], is given as follows. A quantity $c_i$ (local clustering coefficient of node $i$) is first defined to describe how likely $a_{jm} = 1$ for two neighbors $j$ and $m$ of node $i$. It is defined as the ratio between $e_i$ and $k_i(k_i-1)/2$, in which $e_i$ denotes the actual number of edges between the neighbors of node $i$, i.e.,

$$c_i = \frac{2e_i}{k_i(k_i-1)} = \frac{\sum_{j\neq i} a_{ij}a_{jm}a_{mi}}{k_i(k_i-1)}$$

(2)

The clustering coefficient of a graph is the average of $c_i$ over all nodes:

$$C = \frac{1}{N} \sum_{i\in N} c_i$$

(3)

Thus, $0 \leq c_i \leq 1$. The clustering coefficient indicates the local clustering property and shows the fault tolerance characteristic. Taking the subway network as an example, when one track is out of function, the traffic will not be affected if the neighboring stations are connected. Thus, a larger value of $C$ denotes a better tolerance to fault in a local scale.

2.3. Network (Structural) Efficiency

Efficiency $E$, introduced by Latora and Marchiori [13], is a measure of how efficient information is exchanged over the network. Denoted as $e_{ij}$, efficiency of transfer from nodes $i$ to $j$ is taken as being inversely proportional to the shortest path length, i.e., $e_{ij} = \frac{1}{d_{ij}}$, and the network efficiency $E$ is defined as:

$$E(G) = \frac{1}{N(N-1)} \sum_{i\neq j} e_{ij} = \frac{1}{N(N-1)} \sum_{i\neq j} \frac{1}{d_{ij}}$$

(4)

Note that $E(G)$ is the global efficiency of the whole network and is denoted as $E_{glob}$. Also, $E(.)$ can be defined to characterize the local properties of $G$ by evaluating the efficiency of $G$, the subgraph consisting of the neighbors of node $i$ but excluding node $i$. The local efficiency $E_{loc}$ is then defined as the average efficiency of all subgraphs:

$$E_{loc} = \frac{1}{n} \sum_{i\in G} E(G_i)$$

(5)

$E_{loc}$ plays a similar role as $C$, and tells how efficient the communication between the neighbors of $i$ is in the absence of node $i$, reflecting the robustness of local connection when node $i$ is removed.

However, this definition of $E$ is not fully consistent with the subway operation. In subway networks, segments of some lines overlap, thus affecting transportation efficiency. To correct this, if nodes $i$ and its neighbor $j$ are connected by multiple edges, we scale the link connecting the two nodes by a factor $w_{ij}$ and use the scaled link to compute $d_{ij}$, i.e.,

$$w_{ij} = \frac{1}{n}$$

(6)

where $n$ is the number of edges between station $i$ and its neighbor station $j$. Then, $E$ is calculated based on the weighted network structure.

2.4. Average station coverage area and load

In order to evaluate the average distance from a random passenger to a subway station and the average passenger load of a station, we propose two parameters, namely, average station coverage area (ASCA) and average station load (ASL). Here, we define ASCA as as the ratio of subway network area $S_{all}$ and the number of station, i.e., $\text{ASCA} = \frac{S_{all}}{n}$. Thus, ASCA reveals the average area served by a station, or equivalently, the average distance to a subway station for passengers, and the station density. Moreover, ASL is defined as the ratio of average daily passenger flow $P$ and the number of stations, i.e., $\text{ASL} = \frac{P}{N}$, reflecting on the average crowdness of the stations.

3. Statistical results

In this paper, the subway networks in Beijing, Hong Kong, London, Paris and Tokyo are studied. Basic information of these subway networks are listed in Table 1.

Figure 1 shows the shortest path length distribution of the five subways. It is observed that they basically follow
the Γ distribution and Hong Kong has the smallest value of network diameter, which is defined as the maximal shortest path length of a network.

The characteristic path lengths $L$ for the five systems are listed in Table 2, which shows that Tokyo offers the shortest characteristic path length, and Hong Kong and Paris having slightly longer characteristic path lengths.

Clustering coefficients $C$ are calculated and listed in Table 3, from which we can see that the London subway has a relatively bigger $C$, implying a better tolerance to faults in a local scale.

Efficiency $E$, based on weighted edges as explained in Section 2.3, is compared in Table 4. It can be shown from the result that all those five subway networks behave less efficient in the topological level compared to a fully connected network (which has a theoretical efficiency of 1). This is because the number of edges $Q \ll N(N - 1)/2$ and the neighbors of most nodes are isolated from each other. From the values of $E$ we can see that the Tokyo and Hong Kong subways perform better than others in the global scale while the Tokyo and London systems perform better in the local scale.

In this paper, the area served by a subway network is conveniently taken as a rectangle, whose edges are defined by the position of the farthest stations in the four directions. For example, for the Beijing subway shown in Fig. 2, the rectangular boundaries are decided by the farthest stations: Nanshao, Tiangongyuan, Suzhuang and Lucheng. For a fair comparison, the sea areas within the areas covered by the subway are removed for Tokyo and Hong Kong. Furthermore, as Hong Kong is a mountainous city, where only 25% of the defined rectangle is inhabited, we adjust the effective area served by the subway accordingly. Table 5 lists the areas and passengers served by the individual subway systems.

In order to see the station density variation, we divide the area served into five concentric rectangular regions (rectangular boundaries are decided by the farthest stations: Nanshao, Tiangongyuan, Suzhuang and Lucheng). Station names refer to Beijing system.

Table 3: Clustering coefficients of subway networks

<table>
<thead>
<tr>
<th>City</th>
<th>Beijing</th>
<th>HK</th>
<th>London</th>
<th>Paris</th>
<th>Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>$C$</td>
<td>0.0024</td>
<td>0.0059</td>
<td>0.0409</td>
<td>0.0163</td>
<td>0.0285</td>
</tr>
</tbody>
</table>

Table 4: Efficiency of subway networks

<table>
<thead>
<tr>
<th>City</th>
<th>Beijing</th>
<th>HK</th>
<th>London</th>
<th>Paris</th>
<th>Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{\text{glob}}$</td>
<td>0.1012</td>
<td>0.1526</td>
<td>0.1261</td>
<td>0.1143</td>
<td>0.1560</td>
</tr>
<tr>
<td>$E_{\text{loc}}$</td>
<td>0.0024</td>
<td>0.0058</td>
<td>0.0339</td>
<td>0.0146</td>
<td>0.0319</td>
</tr>
</tbody>
</table>

Table 5: Data on areas and passengers served by subways

<table>
<thead>
<tr>
<th>City</th>
<th>Beijing</th>
<th>HK</th>
<th>London</th>
<th>Paris</th>
<th>Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area served (sq. km)</td>
<td>3217.0</td>
<td>177.2</td>
<td>1920.0</td>
<td>347.0</td>
<td>594.3</td>
</tr>
<tr>
<td>Passengers per day (million)</td>
<td>10.876</td>
<td>4.490</td>
<td>8.245</td>
<td>4.130</td>
<td>8.500</td>
</tr>
</tbody>
</table>
gular rings), along the diagonal direction. Fig. 3 shows the ASCA versus the diagonal distance from center (ring). For instance, ASCA at ring = n is the ASCA of the inner area within the nth rectangular ring. Also the values of ASL are shown in Table 6. We see that the Paris subway has relatively small ASCA and ASL, and therefore has higher density and is more convenient for passengers. In addition, stations in Hong Kong, Paris and Tokyo are basically distributed uniformly over the city.

4. Conclusion

The topological structure of five subway networks are studied in terms of the characteristic path length, clustering coefficient, and network efficiency. We propose two parameters, namely, average station coverage area (ASCA) and average station load (ASL), to evaluate the station density and the level of convenience to passengers. Among the five subway networks, the Hong Kong subway has the smallest characteristic distance, and the Tokyo subway has the highest topological efficiency. The London subway has a larger value of clustering and local efficiency, suggesting that it has a better tolerance to fault in a local scale. The Paris subway offers the highest level of convenience to passengers due to the low ASCA and ASL.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant 61322307.

Table 6: Average station load

<table>
<thead>
<tr>
<th>City</th>
<th>Beijing</th>
<th>HK</th>
<th>London</th>
<th>Paris</th>
<th>Tokyo</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASL/day</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(x1000)</td>
<td>39.69</td>
<td>52.82</td>
<td>23.16</td>
<td>14.00</td>
<td>41.46</td>
</tr>
</tbody>
</table>

References

Bus Transport Network in Hong Kong: Scale-free or not?

Tanuja Shanmukhappa, Ivan Wang-Hei Ho, and Chi K. Tse
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong
The Hong Kong Polytechnic University Shenzhen Research Institute, China
Email: tanuja.shanmukhappa@connect.polyu.hk, ivanwh.ho@polyu.edu.hk, encktse@polyu.edu.hk

Abstract—Bus Transport Network (BTN) is one of the major sub-networks in Public Transport Network (PTN) in any city for daily commute of a large number of passengers. In this paper, we analyze the BTN structure in Hong Kong considering the geographical locations of bus stops and their directed connectivity from a complex network perspective. The presence of various geographical and socio-economic constraints in the city give rise to the bus network structure of 4,065 nodes connected with 7,909 edges. 916 routes are operated by five major franchised bus services in Hong Kong. Through introducing and implementing a novel ‘Supernode’ concept in the BTN, which is a collection of closely associated bus stops with respect to their separations and directions, our work evidently shows that the unweighted, directed network under consideration behaves more closely to a scale-free network after applying such concept. Such finding lays the foundation for future studies on BTN design and bus route management for optimal transport efficiency.

1. Introduction

Being one of the most densely populated cities in the world, one third of the total daily public transport in Hong Kong (HK) is accounted by the Bus Transport Network (BTN) according to the latest census statistics [1]. To facilitate the design and management of the bus system for optimal transport efficiency, the understanding of the topological properties, network structures and the connectivity associated with the bus network is of crucial importance. In addition, the evolution of the city has a major influence on the development of the BTN in HK. For example, before the opening of the Cross-Harbour Tunnel connecting Kowloon and the Hong Kong Island in 1972, China Motor Bus Co. (on the Island side) and Kowloon Motor Bus Co. (on Kowloon side) were the major bus operators. Nowadays, Hong Kong has five major franchised bus services operating throughout the city.

In this paper, we aim to analyze the BTN in HK from a complex network perspective. Network analysis started gaining more importance when the Erdos-Renyi method was first proposed for generating random graphs [2]. However, it was evidently recognized later on that the topologies and evolution of real world networks in our daily lives are governed by much more advanced concepts in the field of complex networks. Complex networks can be exploited to describe a wide range of systems from PTN to the Internet, financial systems to social networks, etc. Its applications cover many fields of systems in the real world [3].

There are a number of previous studies on bus network analysis. For example, a detailed study of the bus network structure and its statistical analysis of five different cities in India was provided by Chatterjee et al. in [4]. This paper also discussed a similar concept in brief called the short-distance station pairs, which combines stops that are geographically close to each other (walkable within few meters) without any direct bus connectivity. Zang et al. [5] analyzed the BTN in Beijing based on complex networks defined in the Space-L and P concepts, where Space-L was used for the analysis of topological properties and Space-P for the transfer properties. A weighted complex network analysis of the travel routes in Singapore was done by Soh et al. [6], in which the topological and dynamical properties of the graph structure were considered for analyzing the transport network. This was the first paper that discussed the study of PTN based on geographical properties (e.g., latitude and longitude information). Ferber et al. [7] modeled and empirically analyzed PTN in 14 different cities across the world by a systematic approach under a number of graph representations. A more exhaustive work in the domain was done by Sienkiewicz et al. [8] on the statistical analysis of 22 public transport networks in Poland using various concepts of complex network.

With respect to the network size considered in all of the above-mentioned works, it is evident that Hong Kong gets one of the most densely connected network structures with more than 4,000 nodes, 7,000 edges and 916 routes in total for all the franchised bus services. Our work here aims at analyzing the geographically-constrained bus network in Hong Kong using complex network concepts. The dataset is extracted from the centralized database from the Hong Kong Government [9]. Considering the bus stops and route connectivity among all the operators, a directed graph is generated in the L-space based on the connectivity list in the dataset. The network is verified for the scale-free property as per the Barabasi-Albert model [10]. Initially, the network does not show strong scale-free property with node representation according to the standard graph theory, but with the newly introduced concept of supernode, we found that the structure behaves more like a scale-free network. Such finding will provide insights into the studies and design of bus network and its management. For instance, bus network interpreted with...
the supernode concept can be easily compared and evaluated with other theoretical scale-free networks generated according to the BA model. Hence, better network design can be identified and management decisions can be made for optimal transport efficiency.

2. The Bus Transport Network in Hong Kong

2.1. Graph Structure

The dataset collected from the centralized repository of the Hong Kong Government [9] is used for the generation of the directed, unweighted graph for the BTN in HK. The graph has \( n \) nodes, where each node represents a bus stop, and \( e \) edges which are connected in Space-L (there exists an edge between two nodes if there is a bus route connecting them directly). Hence the graph is a collection of \( n \) nodes and \( e \) edges: \( G = (N, E) \). Where the set \( N = \{N_i, j=1,2,...,n\} \), set \( E = \{E_e, E_e=(N_i,N_j)\} \), \( N_i, N_j \in N \ & k = 1,2,...,e \). In the digraph, \( N_i \) denotes the edge head while \( N_j \) denotes the edge head.

A \( n \times n \) adjacency matrix is used to describe the connectivity in \( G \) with entries \( a_{ij} \). Where \( a_{ij} = 1 \) if there exists a link between \( N_i \) and \( N_j \), and 0 otherwise. The adjacency matrix is then converted to the edge list which is input to the graph generation using the Gephi tool.

In our case, every node is identified by a unique Node_ID (NID), Node_Name (NN), Latitude (Lat) and Longitude (Long) information. The Lat-Long information is based on the ‘WGS84’ datum standard [11]. Fig. 1 shows the complete distribution of nodes (bus stops) in Hong Kong as viewed on Google Earth based on their Lat-Long information.

![Bus-stop distribution in HK based on Lat-Long](image1)

The node locations are fixed based on the geographical coordinates and the connectivity among them is established using the edge list. The Gephi tool is used for generating the complete graph structure as described above and the finalized network is shown in Fig. 2. The digraph has \( n = 4,065 \), \( e = 7,909 \), and 916 bus routes in total.

![The structure of the BTN in HK based on the geographical locations (Lat-Long) of bus stops.](image2)

2.2. Degree (\( \delta \)) of Nodes

For a digraph, the in-degree (\( \delta_{in} \)) is defined as the total number of edge heads injected to a particular node and the out-degree (\( \delta_{out} \)) is defined as the total number of edge tails connected to the node. Hence, the total degree of a node is \( \delta_{total} = \delta_{in} + \delta_{out} \). By measuring the directed in and out degrees of the network, it is found that every node is typically 2-connected in the Hong Kong BTN. Table 1 illustrates the list of nodes with the highest in- and out-degrees. They are regarded as transfer hubs in the BTN.

<table>
<thead>
<tr>
<th>No</th>
<th>Node Name</th>
<th>( \delta_{in} )</th>
<th>Node Name</th>
<th>( \delta_{out} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Western harbour crossing</td>
<td>23</td>
<td>Lantau link toll plaza</td>
<td>25</td>
</tr>
<tr>
<td>2</td>
<td>Lantau link toll plaza</td>
<td>23</td>
<td>Western harbour crossing</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>Tate’s cairn tunnel</td>
<td>20</td>
<td>Tate’s cairn tunnel</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>Old wan chai police station</td>
<td>14</td>
<td>Cross harbour tunnel</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>Cross harbour tunnel</td>
<td>14</td>
<td>Immigration tower</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td>Beacon heights</td>
<td>13</td>
<td>Tai lam tunnel</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>Cheung on bus terminus</td>
<td>12</td>
<td>Huanggang</td>
<td>12</td>
</tr>
<tr>
<td>8</td>
<td>Tuen mun road bus-bus interchange</td>
<td>12</td>
<td>Cheung on bus terminus</td>
<td>11</td>
</tr>
</tbody>
</table>

3. Node Degree Distribution Analysis

Degree distribution corresponds to the probability of finding a node with degree \( k \) or the probability distribution of node degrees over the complete network. Barabasi et al. [7] showed that if the degree distribution of a network follows power law distribution, then the network is scale-free. This is in contrast to some typical random networks with a binomial or Poisson node degree distribution as
proposed by Erdos-Renyi [2]. Mathematically, the power law distribution is defined as

\[ P(k) \propto k^{-\alpha} \quad \text{for } k > k_{\text{min}}, \]

where \(2 < \alpha < 3\) is the Pareto index and \(k\) is the node degree.

### 3.1. Fitting Power Law to Empirical Data

Consider

\[ P(k) = C k^{-\alpha}, \]

where \(C\) is a constant. Apply log on both sides, we have

\[ \log P(k) = -\alpha \log k + \log C. \]

Hence, on log-log scale, the degree distribution of power law follows a straight line with slope \(-\alpha\). From (3), it is evident that (2) is valid only for \(k > k_{\text{min}}\) and \(\alpha > 1\). After normalizing the distribution, we find that \(C = (\alpha - 1)/k_{\text{min}}^{\alpha - 1}\). Hence (2) is reduced to

\[ P(k) = ((\alpha - 1)/k_{\text{min}}) (k/k_{\text{min}})^{-\alpha}. \]  

Clauset et al. showed that the value of \(k_{\text{min}}\), which is the lower bound on power law is estimated using the Kolmogorov-Smirnov (KS) test [12], the value of \(\alpha\), which is the scaling parameter is estimated based on maximum-likelihood estimation (MLE). After calculating \(k_{\text{min}}\) and \(\alpha\), we perform goodness-of-fit tests between the empirical data distribution and the hypothesized power law distribution and compute the corresponding ‘p-value’. Typically, if the value of \(p \geq 0.1\), power law distribution is a plausible hypothesis. By fitting our data to the power law distribution, we found that \(\alpha = 3.5\), \(k_{\text{min}} = 4\) and the \(p\)-value < 0.1, which indicates that the network does not follow power law distribution and hence does not behave as a scale-free network. Fig. 3 shows the in- and out-degree distributions of the BTN on a double logarithmic scale.

![Fig. 3. In-degree and out-degree distributions of the BTN in HK on a log-log scale.](image)

### 3.2. The Supernode Concept in BTN

As discussed in Section 1, the main idea behind the supernode concept is to consider closely associated bus stops in a BTN as one single node in the graph. The concept of supernode is considered important because it is a more accurate way of looking at the network from a passenger’s perspective. Keeping this in mind, a collection of nearby stops is treated as a transfer hub in the PTN. When we look at the geographical positions of bus stops in Fig. 1, it is obvious that sets of bus stops are typically separated by a few meters and are of walkable distance in geographically constrained cities like Hong Kong (Fig. 4 shows several examples). We aim to combine two or more of such nodes together to form a ‘Supernode’ in the BTN according to some conditions.

**Conditions of forming a supernode**

As mentioned in Section 2.1, every node in the network is identified by a unique NID, NN and Lat-Long information. Below we discuss the necessary conditions for defining a supernode.

**Condition 1: Node_Names and their directions of connectivity**

In the Hong Kong BTN, typically if the bus stops have the same name and a different ID, it indicates that the stops are located on either sides of the road, and hence are opposite to each other. Since such nodes are well within walking distance, we combine such nodes together as one supernode.

**Condition 2: Geographic distance between nodes**

Here we calculate the geographic distance \(d_{ij}\) between two nodes and check if \(d_{ij} < d_{th}\), where \(d_{th}\) represents a certain distance threshold. If \(d_{ij}\) is less than \(d_{th}\), then the nodes are combined as a supernode. In our study, the distance threshold is set to be 100 m, which is mostly considered as a walkable distance.

By considering the above conditions, the supernode concept was implemented in the HK BTN, and we verify its scale-free property again after such implementation. For a distance threshold of 100 m, the network structure is re-defined with the new set of in-degree and out-degree, and the corresponding degree distribution is plotted in Figure 5.
Fig. 5 shows the in-degree and out-degree distributions of the HK BTN after the supernode implementation. From the figure, it is more evident that the network structure now closely behaves like a scale-free network with $\alpha = 3.5$, $k_{\min} = 4$, and $p$-value $> 0.1$ for KS test, which confirms that the new data set with supernode implementation satisfies the scale-free property. In addition, the network shows an average node degree of 4 compared to 2 before considering the supernodes, i.e., a node is now typically 4-connected rather than 2-connected.

![In-degree and out-degree distributions of the HK BTN on a log-log scale with the supernode implementation.](image)

Table 2 shows the in-degree and out-degree statistics before and after the supernode implementation in the HK BTN. Specifically, we find that the number of high-degree nodes (e.g., from degree 10 to 20) have significantly increased after re-defining the network structure under the supernode concept. As a result, it causes the network to behave like a scale-free network.

Table 2. In- and out-degree statistics before and after implementing the supernode concept.

<table>
<thead>
<tr>
<th>In-degree (Supernode)</th>
<th>Number of Nodes</th>
<th>Out-degree (Supernode)</th>
<th>Number of Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>2</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>11</td>
<td>6</td>
</tr>
<tr>
<td>10</td>
<td>7</td>
<td>10</td>
<td>6</td>
</tr>
</tbody>
</table>

4. Conclusion

In this paper, we have analyzed the topographical structure of the BTN in Hong Kong and evaluated its scale-free property. With the standardized graph representation, the network initially does not behave as scale-free. However, with the implementation of the proposed supernode concept in the BTN, we found that the network plausibly behaves as a scale-free network. The presence of supernodes in a scale-free bus transport network provides convenient switching points that facilitate efficient routing and hence reduce the average shortest path between any two nodes in the network. It is therefore of practical relevance to study the impact of network structure on traffic performance and to identify key factors and parameters that affect performance of public transportation systems.

Acknowledgement

This work is partially supported by the Early Career Scheme (Project No. 2520714) established under the University Grant Committee of the Hong Kong Special Administrative Region, China; the National Natural Science Foundation of China (Project No. 61401384); and The Hong Kong Polytechnic University (Projects 4-ZZCZ, G-YBK6, and G-YN17).

References

A Fast Method for Finding the Edge to be Added to Minimize Betweenness Centrality of a Specified Vertex

Toshiyuki Namba, Tatsuki Kohno and Norikazu Takahashi

Graduate School of Natural Science and Technology, Okayama University
3–1–1 Tsushima-naka, Kita-ku, Okayama 700–8530, Japan
Email: {namba,kono}@momo.cs.okayama-u.ac.jp, takahashi@cs.okayama-u.ac.jp

Abstract—Betweenness centrality is a measure that represents the importance of each vertex in a graph. From the viewpoint of the robustness against attacks and failures, it is desired that all vertices take similar values of the betweenness centrality. In this paper, a fast algorithm for finding the edge to be added to minimize the betweenness centrality of a specified vertex is developed. The efficiency of the proposed algorithm is confirmed by some experiments on random graphs and scale-free graphs. It is also shown experimentally that the proposed algorithm is useful for improving the robustness of the graph by an edge addition.

1. Introduction

Since the seminal paper by Watts and Strogatz [1], complex networks have attracted a great deal of attention from researchers in physics, engineering, economics, sociology, and so on. When analyzing a network, it is often important to evaluate the importance of each node quantitatively. Various measures of the importance have been proposed in the literature. Among them, this paper focuses on the betweenness centrality [2] which is well known and widely used in the analysis of complex networks.

From the viewpoint of the robustness against attacks and failures, it is desired that all vertices take similar values of the betweenness centrality. A natural way to make a given network more robust is to add a small number of edges to decrease the largest betweenness centrality of the network as much as possible. As an example, let us consider the graph shown in Fig. 1 (a). If we add an edge to connect vertices 2 and 5, the largest betweenness centrality decreases from 8 to 4. However, finding edges to be added is computationally very expensive because, for each of the nonexisting edges, we have to compute the betweenness centrality for all vertices assuming that the edge is added.

In this paper, we consider the problem of finding the edge to be added in order to minimize the betweenness centrality of a specified vertex. This problem is closely related to the improvement of the robustness of a network mentioned above. If the betweenness centrality of the vertex with the largest betweenness centrality can be greatly decreased by an edge addition, it is expected that the largest betweenness centrality of the graph also decreases. We first provide a few theoretical results about the effect of an edge addition on the number and length of shortest paths between any pair of vertices. We next propose an algorithm, which is based on these theoretical results, for solving the above-mentioned problem. We finally examine the efficiency of the proposed algorithm by some experiments on random graphs and scale-free graphs, and show that the proposed algorithm is much faster than a simple method based on Brandes’s algorithm [3].

2. Betweenness Centrality

2.1. Mathematical Expression of Networks

Throughout this paper, a network is expressed as a simple connected undirected graph \( G = (V,E) \), where \( V = \{1,2,\ldots,N\} \) is the vertex set and \( E = \{e_1,e_2,\ldots,e_M\} \) is the edge set. Because \( G \) is simple and undirected, each member of \( E \) is an unordered pair of distinct vertices.

The set of all shortest paths from vertex \( s \) to vertex \( t (\neq s) \) in \( G \) is denoted by \( G_{st} = (V_{st}, E_{st}) \) where \( V_{st} \subseteq V \) is the set of all vertices that appear in the shortest paths, and \( E_{st} \) is the set of all directed edges that appear in the shortest paths. Note that, unlike \( E \), each member of \( E_{st} \) is an ordered pair of distinct vertices. The set of all shortest paths from vertex \( s \) to all other vertices in \( G \) is denoted by \( G_s = (V,E_s) \) where \( E_s \) is the set of all edges that appear in the shortest paths. It is clear that \( E_{st} \) is a subset of \( E_s \).

The number of shortest paths from vertex \( s \) to vertex \( t \) and the length of these paths are denoted by \( \sigma_{st} \) and \( d_{st} \), respectively. Because \( G \) is undirected, \( \sigma_{st} = \sigma_{ts} \) and \( d_{st} = d_{ts} \) for all pairs of \( s \) and \( t (\neq s) \). In the following discussions,
we assume for the sake of convenience that $\sigma_{ss} = 1$ and $d_{ss} = 0$ for all $s \in V$.

2.2. Betweenness Centrality

The betweenness centrality of vertex $i$, which is denoted by $B_i$, is defined by

$$B_i = \sum_{s \neq i} \sum_{t \neq s, t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

(1)

where $\sigma_{st}(i)$ is the number of shortest paths from vertex $s$ to vertex $t$ that pass through vertex $i$. In other words, the betweenness centrality of vertex $i$ is the sum of the ratio of the number of shortest paths from $s$ to $t$ passing through $i$ to the number of all shortest paths from $s$ to $t$ over all pairs of $s$ and $t$. Therefore, a vertex with a high betweenness centrality is a key for many pairs of vertices because all or many of the shortest paths between these two vertices pass through the vertex.

2.3. Brandes’s Algorithm for Computing Betweenness Centrality

A simple method for computing $B_i$ for all $i \in V$ is as follows. First, for $s = 1, 2, \ldots, N$, one finds all shortest paths from vertex $s$ to all other vertices by using, for example, the breadth-first search algorithm. During this process, $d_{st}$ and $\sigma_{st}$ can be found for all pairs of distinct vertices $s$ and $t$, and $\sigma_{st}(i)$ can be found for all triples of distinct vertices $s$, $t$, and $i$. Next, one computes $B_i$ by using (1) for $i = 1, 2, \ldots, N$. The computational complexity of this method is $O(N^3)$, which becomes very large as the number of vertices increases.

The most widely used algorithm for computing the betweenness centrality is the one proposed by Brandes [3]. It is described as follows.

**Algorithm 1 (Brandes’s Algorithm)**

Input: A simple connected undirected graph $G = (V, E)$

Output: $B_1, B_2, \ldots, B_N$

1. Set $s \leftarrow 1$ and $B_i \leftarrow 0$ for $i = 1, 2, \ldots, N$.
2. Find $G_s = (V, E_s)$ by using the breadth-first search. In so doing, find also $d_{st}$ and $\sigma_{st}$ for $i = 1, 2, \ldots, N$.
3. Set $\delta_s(i) \leftarrow 0$ for $i = 1, 2, \ldots, N$.
4. Update the value of $\delta_s(i)$ by

$$\delta_s(i) \leftarrow \sum_{j \in (i) \in E_s} \frac{\sigma_{sj}}{\sigma_{st}} (1 + \delta_s(j))$$

where $i = 1, 2, \ldots, s-1, s+1, s+2, \ldots, N$.
5. Update the value of $B_i$ by

$$B_i \leftarrow B_i + \delta_s(i), \quad i = 1, 2, \ldots, N.$$

6. If $s = N$ then return $B_1, B_2, \ldots, B_N$ and stop. Otherwise set $s \leftarrow s + 1$ and go to Step 3.

Let us consider the computational complexity of this algorithm. Step 2 can be done in $O(M)$ time, where $M$ is the number of edges. Step 4 can be done in $O(M)$ too, because we can update the values of $[\delta_s(i)]_{i=1}^{N}$ while traversing the vertices once in non-increasing order of their distance from $s$ [3]. Therefore, the total computational complexity of Algorithm 1 is $O(NM)$, which is in general lower than $O(N^3)$. In particular, the former is much lower than the latter if $M \ll N^2$, that is, the graph is sparse.

3. Theoretical Analysis of the Effect of One Edge Addition on Betweenness Centrality

Let $G' = (V, E')$ be the graph obtained from a graph $G = (V, E)$ by adding an edge $e = (\alpha, \beta) \not\in E$. The set of all shortest paths from vertex $s$ to vertex $t$ ($t \neq s$) in $G'$ is denoted by the directed graph $G'_s = (V'_s, E'_s)$. The set of all shortest paths from $s$ to all other vertices in $G'$ is denoted by $G'_s = (V, E'_s)$. The number of shortest paths from $s$ to $t$ and the length of these paths are denoted by $\sigma'_{st}$ and $d'_{st}$ respectively. The number of shortest paths from $s$ to $t$ that pass through $i$ is denoted by $\sigma'_{st}(i)$.

In this section, we show that the values of $\sigma'_{st}$, $\sigma'_{st}(i)$ and $d'_{st}$ can be computed from $(\sigma_{pq})_{p,q=1}^{N}$ and $(d_{pq})_{p,q=1}^{N}$. We hereafter assume without loss of generality that

$$d_{st} \leq d_{st}.$$  

(2)

We also assume for convenience that $V_s = V'_s = \{s\}$, $E_{st} = E_s = \emptyset$, $\sigma_{ss} = \sigma'_{ss} = 1$ and $d_{ss} = d'_{st} = 0$ for all $s \in V$.

**Theorem 1** Under the assumption (2), the following statements hold true.

1. If $d_{st} < d_{st}$ and $d_{st} + d_{st} + 1 < d_{st}$ then i) $G'_s$ consists of $G_{st}$, the directed edge ($\alpha, \beta$) and $G_{st}$, ii) $\sigma'_{st} = \sigma_{ss} \sigma'_{st}$, and iii) $d'_{st} = d_{ss} + 1 + d_{jt}$.
2. If $d_{st} < d_{st}$ and $d_{st} + d_{st} + 1 = d_{st}$ then i) $G'_s$ consists of $G_{st}$, $G_{st}$, the directed edge ($\alpha, \beta$), and $G_{st}$, ii) $\sigma'_{st} = \sigma_{st} \sigma'_{st}$, and iii) $d'_{st} = d_{st}$.
3. If $d_{st} = d_{st}$ and $d_{st} + d_{st} + 1 > d_{st}$ then i) $G'_s = G_{st}$, ii) $\sigma'_{st} = \sigma_{st}$, and iii) $d'_{st} = d_{st}$.

**Theorem 2** Under the assumption (2), the following statements hold true.

1. If $d_{st} < d_{st}$ and $d_{st} + d_{st} + 1 < d_{st}$ then

$$\sigma'_{st}(i) = \begin{cases} \sigma_{ss} \sigma_{st} \sigma'_{st}, & \text{if } i \in V_{st}, \\ \sigma_{st} \sigma'_{st}, & \text{if } i \in V_{st}, \\ 0, & \text{if } i \not\in V_{st} \cup V_{st}. \end{cases}$$
2. If $d_{st} < d_{st}'$ and $d_{st} + d_{it} + 1 = d_{st}'$ then

$$\sigma'_{st}(i) = \begin{cases} 
\sigma_{st}(i) + \sigma_{st}\sigma_{st}, & \text{if } i \in V_{sa}, \\
\sigma_{st}(i) + \sigma_{st}\sigma_{st} + d_{st}, & \text{if } i \in V_{bt}, \\
\sigma_{st}(i), & \text{if } i \notin V_{sa} \cup V_{bt}.
\end{cases}$$

3. If $d_{st} = d_{st}'$ or $d_{st} + d_{it} + 1 > d_{st}'$ then $\sigma'_{st}(i) = \sigma_{st}(i)$.

We omit the proofs of these theorems due to space constraints.

Note that we need to check the conditions $i \in V_{sa}$ and $i \in V_{bt}$ in order to compute $\sigma'_{st}(i)$. However, this is easily done because it is well known that $i \in V_{pq}$ if and only if $d_{pq} = d_{pq} + d_{it}$. Note also that the value of $\sigma_{st}(i)$ is required for the computation of $\sigma'_{st}(i)$. However, $\sigma_{st}(i)$ can be easily obtained by

$$\sigma_{st}(i) = \begin{cases} 
\sigma_{st}\sigma_{st}, & \text{if } d_{si} + d_{it} = d_{st}, \\
0, & \text{if } d_{si} + d_{it} \neq d_{st}.
\end{cases}$$

Thus $\sigma'_{st}(i)$ can be computed from $\{\sigma_{pq}\}_{p,q=1}^N$ and $\{d_{pq}\}_{p,q=1}^N$.

4. Proposed Algorithm

We consider in this section the problem of finding the edge to be added to minimize the betweenness centrality of a specified vertex. For this problem, the following algorithm is easily derived from the theoretical results in the previous section.

Algorithm 2

Input: $G = (V, E)$ and $i \in V$.
Output: Edge $(\alpha, \beta) \notin E$ to be added.

1. Compute $\{\sigma_{pq}\}_{p,q=1}^N$ and $\{d_{pq}\}_{p,q=1}^N$ for $G$ by the breadth-first search algorithm.
2. Set $B_{min} \leftarrow \infty$ and $\alpha \leftarrow 1$.
3. Set $\beta \leftarrow 1$.
4. If $(\alpha, \beta) \in E$ then go to Step 7. Otherwise go to Step 5.
5. Compute $B'_i$ by Algorithm 3 given below.
6. If $B'_i < B_{min}$ then set $B_{min} \leftarrow B'_i$, $\alpha_{min} \leftarrow \alpha$ and $\beta_{min} \leftarrow \beta$.
7. If $\beta = N$ then go to Step 8. Otherwise set $\beta \leftarrow \beta + 1$ and go to Step 4.
8. If $\alpha = N - 1$ then return $(\alpha_{min}, B_{min})$ and stop. Otherwise set $\alpha \leftarrow \alpha + 1$ and go to Step 3.

Algorithm 3

Input: $G = (V, E)$, $\{\sigma_{pq}\}_{p,q=1}^N$, $\{d_{pq}\}_{p,q=1}^N$, $(\alpha, \beta) \notin E$ and $i \in V$.
Output: $B'_i$

1. Set $s \leftarrow 1$ and $B'_i \leftarrow 0$.  

<table>
<thead>
<tr>
<th>$N$</th>
<th>$M$</th>
<th>Proposed (sec)</th>
<th>Simple (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>387</td>
<td>0.062</td>
<td>12.453</td>
</tr>
<tr>
<td>100</td>
<td>873</td>
<td>0.265</td>
<td>96.812</td>
</tr>
<tr>
<td>150</td>
<td>1867</td>
<td>1.281</td>
<td>986.25</td>
</tr>
<tr>
<td>200</td>
<td>3317</td>
<td>4.031</td>
<td>5910.172</td>
</tr>
<tr>
<td>250</td>
<td>5083</td>
<td>9.968</td>
<td>23525.422</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$N$</th>
<th>$M$</th>
<th>Proposed (sec)</th>
<th>Simple (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>204</td>
<td>0.094</td>
<td>7.203</td>
</tr>
<tr>
<td>100</td>
<td>294</td>
<td>0.328</td>
<td>40.891</td>
</tr>
<tr>
<td>150</td>
<td>444</td>
<td>1.609</td>
<td>302.077</td>
</tr>
<tr>
<td>200</td>
<td>594</td>
<td>4.968</td>
<td>1197.421</td>
</tr>
<tr>
<td>250</td>
<td>744</td>
<td>11.890</td>
<td>3520.593</td>
</tr>
</tbody>
</table>
Table 3: The largest betweenness centrality before and after an edge addition to random graphs.

<table>
<thead>
<tr>
<th>N</th>
<th>M</th>
<th>Before</th>
<th>After</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>387</td>
<td>83.261</td>
<td>76.778</td>
<td>-6.483</td>
</tr>
<tr>
<td>100</td>
<td>873</td>
<td>110.542</td>
<td>107.266</td>
<td>-3.276</td>
</tr>
<tr>
<td>150</td>
<td>1867</td>
<td>132.943</td>
<td>130.914</td>
<td>-2.029</td>
</tr>
<tr>
<td>200</td>
<td>3317</td>
<td>183.125</td>
<td>181.772</td>
<td>-1.353</td>
</tr>
<tr>
<td>250</td>
<td>5083</td>
<td>243.350</td>
<td>241.127</td>
<td>-2.223</td>
</tr>
</tbody>
</table>

Table 4: The largest betweenness centrality before and after an edge addition to scale-free graphs.

<table>
<thead>
<tr>
<th>N</th>
<th>M</th>
<th>Before</th>
<th>After</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>204</td>
<td>1327.625</td>
<td>1307.093</td>
<td>-20.532</td>
</tr>
<tr>
<td>100</td>
<td>294</td>
<td>3330.202</td>
<td>3303.231</td>
<td>-67.971</td>
</tr>
<tr>
<td>150</td>
<td>444</td>
<td>8228.860</td>
<td>8178.460</td>
<td>-50.400</td>
</tr>
<tr>
<td>200</td>
<td>594</td>
<td>15834.116</td>
<td>15776.048</td>
<td>-58.068</td>
</tr>
<tr>
<td>250</td>
<td>744</td>
<td>25438.355</td>
<td>25367.843</td>
<td>-70.512</td>
</tr>
</tbody>
</table>

of $p$ was set to 0.1666. Scale-free graphs were generated by the process proposed by Barabási and Albert [5]. To be more specific, we start with the complete graph with $m_0$ vertices, and then add new vertices one by one. When a new vertex is added, it is connected to $m (\leq m_0)$ existing vertices with a probability that is proportional to their degrees. In the experiments, the values of $m_0$ and $m$ were set to 4 and 3, respectively. For each of the graphs mentioned above, the vertex with the largest betweenness centrality was chosen as the specified vertex. All algorithms were implemented with C language, compiled with gcc version 5.3.0 and executed on a PC with Intel Core i5 4590 processor and 8GB RAM.

Computation time of the two algorithms is shown in Tables 1 and 2. In the case of random graphs, the proposed algorithm is faster than the simple one by a factor of 200 to 2360 as shown in Table 1. In the case of scale-free graphs, the proposed algorithm is faster than the simple one by a factor of 77 to 296 as shown in Table 2. These results suggest the effectiveness of the proposed algorithm.

If the edge found by the proposed algorithm is added to the graph, the betweenness centrality of the vertex having the largest betweenness centrality among all vertices is minimized. However, it is not guaranteed that the largest betweenness centrality of the graph decreases. If the betweenness centrality of some vertex other than $i$ is the vertex having the largest betweenness centrality before the edge is added, becomes greater than $B_i$ by the addition of the edge, the largest betweenness centrality increases. So let us examine whether or not the largest betweenness centrality is decreased by the edge addition. Results are shown in Tables 3 and 4, from which we see that the largest betweenness centrality decreases for all graphs used in the experiments. However, this is not always true. To see this, let us consider the graph shown in Fig. 2. Among 11 vertices of the graph, vertex 3 has the largest betweenness centrality, which is 11.50. If Algorithm 2 is applied to vertex 3 of this graph, it returns edge [2, 10]. In fact, this edge can decrease the betweenness centrality of vertex 3 to 2.95. However, it also increases the betweenness centrality of vertex 2 from 10.75 to 14.50.

6. Conclusion

The problem of finding the edge to be added that minimizes the betweenness centrality of a specified vertex of a graph has been studied in this paper. We first analyzed theoretically the effect of an edge addition on the number and length of the shortest paths between each pair of vertices. We then developed an algorithm, which is based on the theoretical analysis, for solving the problem and demonstrated the efficiency by experiments.

Acknowledgement

This work was partially supported by JSPS KAKENHI Grant Number JP15K00035.

References

Which vertices affect the spread of disease in temporal networks?

Koshi Abe†, Yutaka Shimada‡, Kantaro Fujiwara‡ and Tohru Ikeguchi‡

†Department of Management Science, Graduate School of Engineering, Tokyo University of Science
‡Department of Information and Computer Technology, Faculty of Engineering, Tokyo University of Science
6-3-1 Niijuku, Katsushika, Tokyo 125-8585, Japan
Email: koshi.abe@hisenkei.net

Abstract—To prevent outbreaks of infectious diseases, it is one of the effective approaches to analyze mathematical models of the infectious diseases. To model the spread of the infectious diseases in realistic situations, contact networks of individuals have attracted much attention in recent years. In this paper, using a simple mathematical model of the infectious disease, we investigate the spread of the infectious disease on the real contact networks observed from person-to-person interactions recorded by radio frequency identifier at a hospital and a high school. We also investigate how removal of vertices with high centralities from networks affect the spread of the infectious diseases. As a result, we find that removal of the vertices with the highest number of contacts is the most effective method for suppressing the spread of the infectious diseases in the hospital. On the other hand, in the high school, the closeness centrality-based removal of vertices is more effective than the strategy based on the number of contacts.

1. Introduction

Several common features of real networks including the Internet, gene networks [1], economic systems, and face-to-face interactions between individuals [2] have been clarified. The small-world [3] and the scale-free [4] properties are known as the universal properties.

In previous research, information diffusion and spread of diseases have been discussed on static networks whose structures do not change with time [5]. However, the structures of real networks always change with time in the realistic situation [6].

We investigate how the infectious disease spreads on real temporal networks whose structures change with time, using the data of person-to-person interactions recorded by the radio frequency identifier (RFID) in a hospital and a high school [7]. In Ref. [8], we have reported that the infectious disease more widely spread on a real hospital network than a real high school network. This result in Ref. [8] is caused by the differences between the structure of the hospital and the high school networks. From this result, we can infer that an effective method for preventing undesirable outbreaks is different between the hospital network and the high school network. In this paper, we investigate how we can effectively slow down the spread of the infectious disease by removing vertices from the temporal networks. We also identify vertices which are most influential in the temporal networks.

As a result, removal of the vertices with the highest number of contacts can effectively suppress the spread of the infectious disease in the hospital network. On the other hand, removal of the vertices with the highest degrees and those with the highest closeness centralities is more effective method for preventing the infectious diseases from diffusing in the high school network than the removal of vertices with the highest number of contacts.

2. Data

In this paper, we used the data of face-to-face contacts of individuals recorded by SocioPatterns [7, 9, 10]. The data are classified into two types. The first data were observed at a hospital in Lyon, France, and the second were observed at a high school in the Lycée Thiers, Marseilles, France. The contacts between individuals were recorded every 20 seconds.

In the data of the hospital, subjects participating in the experiments were 29 patients and health care workers including 27 nurses and nurses’ aides, 11 medical doctors, and 8 administrative staff members. The number of the subjects was 75. The data were collected over five days (see Ref. [9] for details). The number of contacts in the hospital was 32,424 during the five days.

In the data of the high school, the subjects were 126 high school students. The data were collected over four days (see Ref. [10] for details). The number of contacts in the high school was 28,561 during the four days.

3. Methods

We investigate which vertices are influential in the diffusion of the infectious disease on temporal networks by numerical simulations. We first constructed networks from the real data and then applied a mathematical model of the infectious disease to these networks. The edge between the vertex \(i\) and the vertex \(j\) at time \(t\) is described by \(l_{ij}(t) \in [0, 1]\), where if a contact exists, \(l_{ij}(t) = 1\), otherwise \(l_{ij}(t) = 0\). A state of the vertex \(i\) at time \(t\) is described by \(S_i(t) \in \{0, 1\}\), where if the vertex \(i\) is infected with the infectious disease, \(S_i(t) = 1\), otherwise \(S_i(t) = 0\).
In this experiment, we first choose the vertex \( i \) randomly and change its state into an infected state \( S_i(t_1) = 1 \), where \( t_1 \) is the time when the vertex \( i \) contacts the other vertices at the beginning in these data. We also define an infection rate \( r \) as the probability that a state of a susceptible vertex changes to an infected state. The number of infected subjects at time \( t \) is \( I(t) (t = 0, 20, \cdots , T) \), where \( T \) is the time when the infectious disease disappears from the networks. The ratio of the number of infected subjects at time \( t \) to the total number of subjects is \( P(t) = I(t)/N \), where \( N \) is the number of subjects. When the vertex \( i \) contacts with the vertex \( j \) under the situation where \( S_i(t) = 1 \) and \( S_j(t) = 0 \) at time \( t \), the infectious disease is transmitted from the vertex \( i \) to the vertex \( j \) with the probability \( rl_{ij}(t) \). When the vertex \( j \) is infected at time \( t \) (\( S_j(t) = 1 \)), the vertex \( j \) has the infectious disease during a fixed period \( \tau \). After the period \( \tau \), the state of the vertex \( j \) changes to the recovered state. Recovered vertices are not infected again, and they do not transmit the infectious disease to other vertices.

We first investigate how the infectious disease spreads on the contact networks by \( P(t) \). We next try to suppress the spread of the infectious disease by removing a few vertices from the networks. The number of removed vertices is defined by \( R \). We removed \( R \) vertices based on their characteristic features: degrees, the number of contacts, the closeness centrality, the betweenness centrality, and the eigenvector centrality [11]. We calculate these centralities from the temporally aggregated static networks by using all contact data. The closeness centrality \( c_i \) of the vertex \( i \) is defined as follows:

\[
c_i = \frac{N - 1}{\sum_{j \neq i}^N d(v_i, v_j)},
\]

where \( d(v_i, v_j) \) is the shortest path length between the vertices \( i \) and \( j \). The betweenness centrality \( b_i \) of the vertex \( i \) is defined as follows:

\[
b_i = \frac{\sum_{i \neq j \neq k}^N g_{ijk}^{(i,j,k)}}{(N - 1)(N - 2)/2},
\]

where \( N_{i,j,k} \) is the total number of the shortest paths between the vertices \( i, j \), and \( g_{ijk}^{(i,j,k)} \) is the number of the shortest paths passing through the vertex \( i \) out of all the shortest paths between the vertices \( i, j \). The eigenvector centrality of the vertex \( i \) is described as \( u_i \), and \( u \equiv (u_1, u_2, \cdots , u_N)^T \). Let \( A \) be an adjacency matrix of a network. Let \( \lambda_n \) be the nth eigenvalue of \( A \). Then, the eigenvector centrality is given by

\[
\lambda_i u = \lambda u_i,
\]

that is, the vector of the eigenvector centralities is the eigenvector of \( A \) corresponding to the maximum eigenvalue \( \lambda_1 \).

4. Results

4.1. The relation between the recovery time \( \tau \) and the normalized number of infected subjects \( P(T) \)

We first conducted experiments under the condition that \( r = 0.01 \). We calculated the final ratio of the number of infected subjects to the total number of subjects \( P(T) \). Figure 1 shows how \( P(T) \) changes when the recovery time \( \tau \) changes. From Fig. 1, \( P(T) \) in the case of the hospital is higher than that of the high school in all values of the recovery time \( \tau \). We supposed that there are two causes of the different tendency of the spread of the infectious disease between the hospital and the high school. One is the difference between the number of contacts in the hospital and that in the high school. The other is the difference between structural properties in the hospital and that in the high school. In Ref. [8], we have investigated how the infectious disease spreads on temporal networks. As a result, the infectious disease easily spreads when the number of contacts is large, and the structural properties affect the diffusion of the infectious disease. From these results, we infer that the properties of the influential vertices in the hospital network and in the high school network are different from each other.

![Figure 1: The relation between the recovery time \( \tau \) and the normalized number of infected subjects \( P(T) \).](image)

The range of \( \tau \) is \( 3 \leq \tau \leq 90[h] \). The infection rate \( r \) is 0.01. The results are averaged over 1,000 trials.

4.2. Preventing the spread of the infectious disease by the removal of vertices

We here try to prevent the infectious disease from widely spreading over networks by removing vertices based on the centrality measures. The degree, the number of contacts, the closeness centrality, the betweenness centrality, and the eigenvector centrality of each vertex are calculated. Then, we remove \( R \) vertices based on these centrality measures, that is, we remove \( R \) vertices with the highest values of these centrality measures.

Figures 2 and 3 show how \( P(T) \) changes when the recovery time \( \tau \) changes at the hospital and the high school. In Figs. 2 and 3, we remove \( R \) vertices with the highest values of centralities which are the degree (red lines),
the number of contacts (blue lines), the closeness centrality (yellow lines), the betweenness centrality (purple lines), and the eigenvector centrality (green lines). The black lines show the results for which vertices are randomly removed. From Fig. 2(a), (b), and (d), removal of the vertices with the highest number of contacts achieve the lowest values of $P(T)$ in all values of $\tau$. On the other hand, from Fig. 2(c), the results are almost the same as each other. This is because the number of removed vertices is too small to slow down the spread of the infectious disease in spite of the high infection rate $r$. From these results, in the hospital, removing vertices with the highest number of contacts is the most effective method for slowing down the spread of the infectious disease.

Figure 2: The relation between the recovery time $\tau$ and the ratio of the number of infected subjects $P(T)$ at time $T$, in the case of the hospital. The results are averaged over 1,000 trials.

Figure 3 shows how $P(T)$ changes when the recovery time $\tau$ changes at the high school. Figure 3(a) and (c) is the result of removing five vertices ($R = 5$). Figure 3(b) and (d) is the result of removing ten vertices ($R = 10$). From Fig. 3, removing the vertices with the high degrees, with the high closeness centrality, and with the high eigenvector centrality leads to the good results that we can effectively slow down the diffusion of the infectious disease. In contrast to the hospital, removal of the vertices with the highest number of contacts is not so effective. Then, removing vertices with the high closeness centrality or those with the high degree are the effective method for slowing down the spread of the infectious disease in the high school.

Figure 3: The relation between the recovery time $\tau$ and the ratio of the number of infected subjects $P(T)$ at time $T$, in the case of the high school. The results are averaged over 1,000 trials.
5. Conclusion

We constructed temporal networks by using the data of person-to-person interactions recorded by RFID [2]. We investigated how the infectious disease spreads on the temporal networks by using a simple mathematical model of the infectious disease on the temporal networks. As a result, we found that the infectious disease spreads more widely in the hospital than in the high school. From this result, we conjectured that the effective methods for slowing down the spread of the infectious disease at the hospital are different from those at the high school.

We then investigated which vertices mainly affect the spread of the infectious disease. As a result, the vertices with the highest number of contacts significantly accelerate the spread of the infectious disease in the hospital, but does not in the high school. In contrast to the hospital, the vertices with the highest closeness centralities and with the highest degrees mainly contribute to the spread of infectious disease in the high school.

Acknowledgment

The authors would like to thank Mr. S. Ishii and AGS Corp. for their kind encouragements on this research. The research of Y.S. is supported by Grant-in-Aid for Research Activity Start-up (No. 26880020) and by Grant-in-Aid for Young Scientist B (16K16126) from JSPS. The research of K.F. was partially supported by Grant-in-Aid for Young Scientist B (16K16138) and by Grant-in-Aid for Challenging Exploratory Research (No 15K12137) from JSPS. The research of T.I. was partially supported by Grant-in-Aid for Exploratory Research (No. 24650116) from JSPS.

References

A Hardware Cellular-Automaton Architecture for Spatial Pattern Generation towards Motion-Vector Estimation of Textureless Objects

Aoi Tanibata†, Miho Ushida†, Alexandre Schmid‡, Masayuki Ikebe†, Tetsuya Asai†, Masato Motomura†

†Graduate School of Information Science and Technology, Hokkaido University
Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan
‡Microelectronic Systems Laboratory, École Polytechnique Fédérale de Lausanne
CH-1015, Lausanne, Switzerland
Email: tanibata@lalsie.ist.hokudai.ac.jp

Abstract

In [1, 2], we proposed a hardware-oriented cellular-automaton algorithm that generates the spatial patterns of textureless objects and backgrounds to estimate the motion-vectors of textureless moving objects. In this report, we propose the fundamental architecture for the algorithm. The system consists of one-dimensional shift-register arrays and arithmetic operators for diffusion, a lookup table acting as a nonlinear “reaction” function, and several state controllers (counters, multiplexers, etc.), which together act as a one-dimensional reaction-diffusion streaming processor (RDSP). Two-dimensional image processing is performed for texture generation by arranging the one-dimensional RDSP on a two-dimensional mesh in a time-division manner, which reduces the complexity of the circuit system as a whole.

1. Introduction

Motion estimation is used in various applications such as hand gesture user interfaces[3] and automatic anomaly detection in monitoring camera pictures[4]. This technique has been actively researched in recent years. The block-matching method is one of the most common methods for motion estimation. However, when this method is applied to textureless moving objects, it can detect the motion vector of the outline only (Fig.1 (a)). In this case, we cannot perceive it as a textureless object or a frame. To estimate the motion more precisely, we have to detect the motion vectors in the outline of textureless moving objects. Therefore, we propose a hardware-oriented cellular-automaton algorithm that generates the spatial patterns of textureless objects and backgrounds in order to estimate the motion vector of textureless moving objects[1, 2]. The textureless moving objects are regarded as objects with the same patterns as the generated spatial patterns. Accordingly, we can detect the motion vectors in the outline of the textureless moving objects (Fig.1 (b)). In this study, we propose a fundamental module and architecture for this algorithm.

2. Algorithm

We proposed an algorithm that generates the spatial patterns using the reaction-diffusion (RD) model[1, 2]. The RD model is a well-known method for spatial pattern generation[5]. One-dimensional RD is obtained by iterative updating. This updating consists of three processes: diffusion, subtraction, and amplification. The diffusion process involves iterative blurring. Blur is described as

$$a_i(t+1) = \frac{a_{i-1}(t) + 2a_i(t) + a_{i+1}(t)}{4}$$

where $i$ is the $i$th pixel in a row of an image, and $t$ is $i$ times blurring. The subtraction process involves finding the difference between the before diffusion and after it. The amplification process amplifies the value using the sigmoid function. In addition, we have to reduce the adverse influence of noise. Therefore, we add a filter that updates after every iteration of updating. We perform only diffusion in filter updating because smoother spatial patterns can be generated through this process.

In two-dimensional RD, the two-dimensional input image is first divided into a one-dimensional arrangement, $x$ and $y$. These arrangements are then repeatedly processed by the one-dimensional RD model. Finally, they are multiplied together [1, 2].
3. Circuits

3.1. Module for One-dimensional Reaction-Diffusion

The proposed module for one-dimensional RD is shown in Fig. 2. In addition, the state transition diagram of the module is shown in Fig. 3. The one-dimensional RD algorithm is considered as switching and repetition of two operations: blurring, and subtraction and amplification. Therefore, we prepare the module as a state machine in this study as shown in Fig. 3.

In Fig. 2, \( i \) is the number of blurrings and \( u \) is the number of updates. In this case, diffusion is obtained by blurring \( i \) times. A one-dimensional RD is obtained by updating \( k \) times. Normal refers to the normal updating consisting of three processes including diffusion, subtraction, and amplification. Filter refers to filter updating.

We also devised a state controller consisting of counters. These counters count the number of blurrings, updates, and pixels in one row of an image. The multiplexers and control signals for the first-in first-out memories (FIFO) consisting of shift registers are controlled by the signals from the controller, and the module can be switched between states.

The module goes into the blurring state during diffusion and filtering. It transmits a pixel value as the output from the imager to the shift-register. It obtains the necessary pixel values from the shift-register. The module loads from the last blurring result instead of the imager from the second iteration onward. It saves the pixel values loaded as the values obtained before diffusion in the FIFO if blurring occurs immediately after updating.

The module goes into the subtraction and amplification state during subtraction and amplification. It subtracts the values after diffusion from the values before diffusion and amplifies the result using the sigmoid function.

3.2. Architecture for Two-dimensional Reaction-Diffusion

The proposed architecture for two-dimensional RD is shown in Fig. 4. Dif1d is the module used for one-dimensional RD.

We also prepared a state controller consisting of some counters similar to the one used in one-dimensional RD. The controller controls the signals for static random access memories (SRAM) and the direction of readout pixel values, in addition to the signal for the one-dimensional RD module. We have to maintain the initial value in two-dimensional RD. The initial values are divided into a one-dimensional arrangement, \( x \) and \( y \) in rotation. Therefore, the architecture begins by saving the pixel values loaded from the imager in the SRAM for input. At the same time, it processes \( x \) though the module for one-dimensional RD and saves the result in SRAM for output. After this, the architecture obtains the initial values in \( y \) from the SRAM for input and processes in \( y \). Finally, it multiplies the results of one-dimensional RD of \( x \) and \( y \) and provides it as the results of two-dimensional RD.

Two-dimensional image processing for texture generation is performed by arranging the module for one-dimensional RD on a two-dimensional mesh in a time-division manner, which reduces the complexity of the circuit system as a whole.

4. Results

4.1. One-dimensional reaction-diffusion

The bit width of the pixel values affects the precision of the result of one-dimensional RD in terms of module one-dimensional RD. We process data with widths of 12 bit and 8 bit through one-dimensional RD and determine the Fast fourier transform (FFT) of the result.

In this case, diffusion is achieved by blurring 25 times. A one-dimensional RD is obtained by updating 10 times. We add a filter updating after every 4 iterations of updating. The gain of the sigmoid function is 5.
Figure 5: Comparison results of 12-bit and 8-bit in one-dimensional reaction-diffusion 1: (a)(d) initial values, (b)(e) result of one-dimensional reaction-diffusion, (c)(f) result of FFT.

The simulation results are shown in Fig.5. The waveform of the 12-bit data is smoother than that of the 8-bit data. The waveform of 8-bit data is sufficiently smooth at a glance. However, we have to consider the result of the FFT. In the result of the 12-bit data, the necessary frequency component has more digits than the noise component. On the other hand, the necessary frequency component in the result of the 8-bit data has approximately the same number of digits as that in the noise component. In this case, we can reduce the noise easily in 12-bit data, but we cannot do so in the 8-bit data. In Fig. 6, we repeat the simulation by changing the initial values and obtain the same result. Therefore, we need a width of at least 12-bit for the pixel values.

4.2. Two-dimensional reaction-diffusion

The generated spatial pattern of the two-dimensional input image is shown in Fig.7. In this case, we used a picture of 120*120 pixels. We assume that we detect the motion vectors of textureless moving objects and prepare a picture of textureless objects against a real background.

The spatial patterns are generated for textureless objects and backgrounds. The same spatial patterns are generated between two frames for textureless objects, regardless of the changing position of textureless objects. Therefore, it appears that the detected motion vectors are in the outline of the textureless moving objects. In Fig. 8, we repeat the simulation by changing the backgrounds and obtain the same result.

Figure 6: Comparison results of 12-bit and 8-bit in one-dimensional reaction-diffusion 2: (a)(d) initial values, (b)(e) result of one-dimensional reaction-diffusion, (c)(f) result of FFT.

Figure 7: Result of two-dimensional reaction-diffusion with a real background.
5. Summary

In this study, a module is used as the state machine in one-dimensional RD and arranged on a two-dimensional mesh in a time-division manner in two-dimensional RD. Therefore, we can reduce the complexity of the circuit system as a whole. In addition, we determine the appropriate bit width of pixel values for one-dimensional RD. In future, we seek to develop a hardware implementation that enables real-time processing.

Acknowledgments

This study was supported by the JSPS Grants-in-Aid for JSPS Fellows, and a Grant-in-Aid for Scientific Research on Innovative Areas [2511001503] from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan.

References


A Probabilistic Model of Nano-Carbon Materials Based on Probabilistic Current

Shohei Hamada†, Aiko Kawane †, Hisato Fujisaka †, and Takeshi Kamio†

†Faculty of Information Science, Hiroshima City University
Ozuka-higashi, Asa-minami-ku, Hiroshima, 731-3194 Japan
Email: hs199296@sp.info.hiroshima-cu.ac.jp

Abstract—We derived an equation of continuity for the probability density function of electrons on nanocarbon materials. From the derived equation, we built a cellular array model of the materials with hexagonal lattice structure. A virtual electron particle moves on the lattice probabilistically along with probabilistic current in the equation of continuity. From numerical experiments, we found that the cellular array successfully provided numerical samples of electron trajectories on the materials.

1. Introduction

Nanocarbon materials like graphene [1] and carbon nano tubes (CNT) [2] are expected to be used for quantum effect devices because of their high electron mobility and their intrinsic quantum mechanical properties such as Klein tunneling.

Quantum effect devices are often represented by probabilistic models in circuit simulators. If the behavior of electrons in the devices is described by a scalar type of Schrödinger equation, Nelson’s stochastic quantization theory is often applied to build the models [3]. However, it is difficult to apply the theory into modeling nanocarbon-based quantum effect devices because electron behavior in the nanocarbon materials is described by a different type of equation, two-dimensional massless Dirac equation [4]. A new method must be established to build the models of nanocarbon-based devices.

In this paper, an equation of continuity for the probability density of an electron in the nanocarbon material is derived. Then, based on the derived equation, a cellular array model of the materials with hexagonal lattice structure is built as a kind of probabilistic cellular automata [5].

2. Nanocarbon Materials

Graphene sheets are composed of carbon atoms bonding to one another as shown in Fig. 1. The carbon atoms A and B, lattice points of a hexagonal lattice, are non-equivalent and adjacent two carbon atoms A and B form a unit cell of graphenes. The behavior of electrons on graphenes is described by the following approximate equation:

\[
\frac{i\hbar}{\partial t} \Psi(x, y, t) = \left[ \hbar v_F \left( \sigma_x \hat{k}_x + \sigma_y \hat{k}_y \right) + V(x, y) \right] \Psi(x, y, t) \tag{1}
\]

where \( \hbar \) is the Plank constant divided by \( 2\pi \), \( v_F \) is the Fermi velocity, and \( V(x, y) \) denotes static potential distribution. Wave function \( \Psi(x, y, t) \),

\[
\Psi(x, y, t) = \begin{pmatrix} \psi_A(x, y, t) \\ \psi_B(x, y, t) \end{pmatrix}, \quad \psi_{A,B} : R^3 \rightarrow C^1 \tag{2}
\]

possesses two elements \( \psi_{A,B}(x, y, t) \) which are wave functions of the electrons on two sublattices consisting respectively of carbon atoms A and B. The operators in Eq. (1) are defined as \( \hat{k}_x \equiv -i\partial/\partial x, \hat{k}_y \equiv -i\partial/\partial y \). Pauli spin matrices \( \sigma_x, \sigma_y \) in the equation are given by

\[
\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} \tag{3}
\]

Equation (1) is the two-dimensional Dirac equation with mass of zero and light speed \( c \) replaced by \( v_F \).

A CNT is considered as a cylindrical graphene sheet in structure. We define two linearly independent vectors \( e_1 \) and \( e_2 \) given by

\[
e_{1/2} = \frac{a}{2}(\sqrt{3}, \pm 1), \quad a = \sqrt{3}a_{AB} \tag{4}
\]
on the graphene sheet, where \( a_{AB} \) is the distance between adjacent two A, B carbon atoms. Let a vector oriented in the circumferential direction of the CNT and a vector parallel to the cylinder axis of the CNT or perpendicular to \( C_h \).
be denoted respectively by $C_h$ and $T$. They are shown in Fig. 2. We express $C_h$ with $e_1$ and $e_2$ as

$$C_h = (C_{h,x}, C_{h,y}) = C_1 e_1 + C_2 e_2 \tag{5}$$

Length of $C_h$ is given by

$$|C_h| = \sqrt{C_1^2 + C_2^2 + C_1 C_2} \tag{6}$$

We set the length as the circumferential length of the CNT. Then, the diameter of the CNT is

$$d_r = \frac{|C_h|}{\pi} \tag{7}$$

In this paper, we assume that $|C_h| \gg a_{AB}$. We set vector $T$ as

$$T = t_1 e_1 + t_2 e_2 \tag{8}$$

$$t_1 = \frac{2C_1 + C_2}{d_r}, \quad t_2 = \frac{C_1 + 2C_2}{d_r}$$

Then, the length of $T$ is

$$|T| = \frac{\sqrt{3}|C_h|}{d_r} \tag{9}$$

If the vectors $C_h$ and $T$ take particular directions, the wave function of electrons on the CNT can be a solution of Eq. (1) with the periodic condition

$$\Psi(x, y, t) = \Psi(x + mC_{h,x}, y + nC_{h,y}, t), \quad m, n: \text{integers} \tag{10}$$

being satisfied.

3. Equation of Continuity

Let the probability density function and the pseudospin density of an electron on a graphene or a CNT be denoted by $\rho(x, y, t)$ and $s_i(x, y, t)$. In addition, we introduce another function $s_i(x, y, t)$. They are given by

$$\rho(x, y, t) = \psi_A(x, y, t)\psi_A^*(x, y, t) + \psi_B(x, y, t)\psi_B^*(x, y, t) \tag{11}$$

$$s_i(x, y, t) = 2\text{Re}(\psi_A^*(x, y, t)\psi_B(x, y, t)) \tag{12}$$

and have the following relations:

$$|\psi_A(x, y, t) \pm \psi_B(x, y, t)|^2 = \rho(x, y, t) \pm s_i(x, y, t) \tag{14}$$

$$|\psi_A(x, y, t) \mp i\psi_B(x, y, t)|^2 = \rho(x, y, t) \mp s_i(x, y, t) \tag{15}$$

Let the complex conjugates of $\Psi$, $\sigma_{\psi y}$, and $\hat{k}_{\psi y}$ be denoted by $\Psi^*$, $\sigma_{\psi y}^*$, and $\hat{k}_{\psi y}^*$, respectively. Then, the complex conjugate equation of Eq. (1) is given by

$$-i\hbar \frac{\partial}{\partial t}\Psi^*(x, y, t) = \left[\hbar v_F \left(\sigma_{\psi y}^* k_x^y + \sigma_{\psi y} k_x^{y*}\right) + V(x, y)\right] \Psi^*(x, y, t) \tag{16}$$

Adding the inner products between Eq. (1) and $\Psi^*$ and between Eq. (16) and $\Psi$, we obtain

$$\frac{\partial}{\partial t}(\psi_A(x, y, t)\psi_A^*(x, y, t) + \psi_B(x, y, t)\psi_B^*(x, y, t))$$

$$= -v_F \left(\frac{\partial}{\partial x} - \frac{\partial}{\partial y}\right)\left((-\psi_A^*(x, y, t)\psi_B(x, y, t) + \psi_A(x, y, t)\psi_B^*(x, y, t))\right)$$

$$= -i(\psi_A^*(x, y, t)\psi_B(x, y, t) - \psi_A(x, y, t)\psi_B^*(x, y, t))$$

Equation (17) can be expressed as

$$\frac{\partial}{\partial t}\rho(x, y, t) + \text{div} J = 0 \tag{18}$$

$$J = v_F (s_x, s_y)^T \tag{19}$$

Since Eq. (18) is so called an equation of continuity, $J$ is considered as probabilistic current. When the wave function of Eq. (1) is a plane wave give by

$$\Phi(x, y, t) = \left(\frac{1}{\sqrt{2\pi}}\right) \exp(i(k_x x + k_y y)) \exp\left(-\frac{E(k_x, k_y)}{\hbar} t\right) \tag{20}$$

$$E(k_x, k_y) = \sqrt{k_x^2 + k_y^2}, \quad \theta = \arctan\left(\frac{k_y}{k_x}\right) \tag{21}$$

the directions of $J$ is given by $\theta$, which coincides with the direction of the propagation of $\Phi$. Then, it is inevitable that Eq. (18) has been derived.

4. Probabilistic Model

A model to be proposed is a hexagonal lattice whose sites A and B are distinguished, as shown in Fig. 1. A virtual electron particle moves from a cite to one of three adjacent sites randomly in a unit time $\Delta t$. We explain the random motion in detail. We denote three vectors from a site A to its three adjacent sites B by $a_i, i = 1, 2, 3$, as shown in Fig. 3, and probabilities that a particle on the site A moves to the three sites B are denoted by $p_i$. The three vectors are given by

$$a_1 = e_1 - \frac{1}{3}(e_1 + e_2) \tag{22}$$

$$a_2 = e_2 - \frac{1}{3}(e_1 + e_2) \tag{23}$$

$$a_3 = -(a_1 + a_2) \tag{24}$$
When the particle does not stay in an identical site for more than one unit time, we have
\[ \sum_{i=1}^{3} p_i = 1 \] (25)

From the discussion at the end of Section 3, a wave packet with momentum distributed around \((k_x, k_y)\) also propagates in the direction of \(\theta\) approximately. Then, the expectation of the velocity of the particle should be proportional to the probabilistic current, which is represented by
\[ \sum_{i=1}^{3} p_i a_i = bJ. \] (26)

Equations (25) and (26) are combined as
\[
\begin{pmatrix}
\frac{1}{b v_{y} s_{x}} \\
\frac{1}{b v_{y} s_{x}}
\end{pmatrix}
= \begin{pmatrix}
\frac{1}{a_{1,x}} & \frac{1}{a_{2,x}} & \frac{1}{a_{3,x}} \\
\frac{1}{a_{1,y}} & \frac{1}{a_{2,y}} & \frac{1}{a_{3,y}}
\end{pmatrix}
\begin{pmatrix}
p_1 \\
p_2 \\
p_3
\end{pmatrix}
\] (27)

where \((a_{1,i}, a_{2,i}, a_{3,i})^T = a_i, i = 1, 2, 3.\) From Eq. (27), probabilities \(p_i\) are obtained as
\[
\begin{pmatrix}
p_1 \\
p_2 \\
p_3
\end{pmatrix}
= \frac{1}{3A} \begin{pmatrix}
1 + 2a_1 \cdot b J \\
1 + 2a_2 \cdot b J \\
1 + 2a_3 \cdot b J
\end{pmatrix}
\] (28)

\[ \Delta = a_1 \times a_2 + a_2 \times a_3 + a_3 \times a_1 \] (29)

where \( \cdot \) and \( \times \) are inner and outer product operators. Coefficient \(b\) should be chosen so that \(p_i \geq 0.\)

By solving Eq. (1), computing Eq. (19) with the solution \( \Psi(x, y, t) \), and using Eq. (28), we determine \(p_i\). Probabilities that the particle moves from a site A to the three adjacent sites B are determined similarly. Then, sample trajectories of the random motion of the particle can be obtained.

5. Numerical Experiments

Let \((x', y')\) be a two dimensional orthogonal coordinate system with axes parallel to \(C_b\) and \(T\). Suppose that the new coordinate system is obtained by rotating \((x, y)\) coordinate system by \(\varphi\). The wave function \(\Psi'(x', y', t)\) of Eq. (1) with coordinate \((x, y)\) replaced by \((x', y')\) has the following relation with \(\Psi(x, y, t)\):
\[
\Psi'(x', y', t) = \left\{ \begin{array}{ll}
\exp\left(\frac{i}{\hbar}k_y y'\right)\Psi_A \\
\exp\left(-\frac{i}{\hbar}k_y y'\right)\Psi_B
\end{array} \right.
\] (30)

From Eqs. (11), (12), and (30), we see that the probability density function and the pseudospin density are conserved on \((x', y')\) plane. Then, we may denote both coordinate systems on graphene and CNT by the same denotation \((x, y)\).

We will determine a wave function in the form of wave packet. The initial conditions of the wave packet are set as follows: Center position: \((x_0, y_0)\), Kinetic momenta: \(h(k_{x0}, k_{y0})\), Variances of the momenta: \(\sigma_{k_x} = \sigma_{k_y} = \sigma_k\). The initial form of the packet can then be expressed as
\[
\phi_0(k_i, k_j) = \frac{1}{\sqrt{2\pi\sigma_k}} \exp\left(-\frac{1}{4}\frac{(k_i - k_{x0})^2}{\sigma_k}\right) \exp\left(-\frac{1}{4}\frac{(k_j - k_{y0})^2}{\sigma_k}\right)
\] (31)

The evolving wave packet is expanded in a series of plane wave solutions (20). Its continuation form is given by
\[
\Psi(x, y, t) = \int \int \Phi(x, y, t) \phi_0 dk_x dk_y
\] (32)

If the integration on \(k_i\) is discretized with step size of \(\Delta k_y\) given by
\[
\Delta k_y = \frac{2\pi}{|C_b|}
\] (33)

that is, Eq. (32) is in the form of the Riemann sum, the wave packet satisfies periodic condition (10). Then, a packet propagating on a CNT is obtained.

Using wave function (32), we compute probabilistic current \(J\) from Eqs. (12), (13), and (19) and then probabilities (28) are determined. Figure 4 shows a wave packet contour-plotted at \(t = 0\) and 30 and a sample trajectory of a virtual electron particle between time interval [0, 30]. They are computed on the following conditions: \(a_{AB} = 0.5, h_{k_{x0}} = 10 \cdot \cos(\pi/4), h_{k_{y0}} = 10 \cdot \sin(\pi/4), \sigma_k = 0.2,\) and \(\Delta t = 0.4.\) Figure 5 shows another pair of a wave packet and a sample trajectory computed on the other conditions that \(a_{AB} = 1.0, h_{k_{x0}} = 10 \cdot \cos(\pi/3), h_{k_{y0}} = 10 \cdot \sin(\pi/3), \sigma_k = 0.2,\) and \(\Delta t = 0.6.\) We see that both the wave packet and the virtual electron particle move almost the same distances in almost the same directions on each condition set.

6. Conclusions

We have presented a method to compute sample trajectories of electrons on nanocarbon materials. Our future subjects include developing the method so that the trajectories distribute in accordance with the probability density function.
Figure 4: A wave packet at $t = 0$ and 30 and a sample trajectory for time interval $[0, 30]$ when $\theta = \pi/4$.

Figure 5: A wave packet at $t = 0$ and 30 and a sample trajectory for time interval $[0, 30]$ when $\theta = \pi/3$.

Acknowledgment

This work was supported by JSPS KAKENHI Grant No. 15K06075.

References


Abstract—This paper considers super-stabilization of
periodic spike-train from the digital spiking neuron consist-
ing of two shift registers connected by a wiring. Depen-
ding on the wiring, the neuron can generate various peri-
odic spike-trains. In order to super-stabilize a desired peri-
odic spike-train, we present a simple deterministic rewiring
method. The super-stabilized periodic spike-trains are ap-
plicable to robust and reliable encoders in multiplexing
communication systems.

1. Introduction

The digital spiking neuron (DSN) is a kind of digital dy-
namical system inspired by integrate-and-fire neuron mod-
els [1]-[3]. The DSN is constructed by two shift registers
connected by a wiring. Depending on the wiring pattern,
the DSN can generate various periodic spike-trains (PSTs)
and transient super-trains to the PSTs. The DSNs are ap-
plicable to spike-based communication systems and spike-
base learning systems [1]-[7].

This paper considers super-stabilization of a desired PST
in the DSN. The super-stability means that almost all initial
points fall directly (instantaneously) into the PST [3]. The
super-stabilized PST is applicable to a robust and reliable
encoder in multiplexing communication systems. In order
to super-stabilize a PST, we introduce a super-stabilizing
wiring method (SSWM [3]). The SSWM is a simple deter-
niministic method that re-wires connection between two shift
registers of the DSN and can super-stabilize a desired PST.

The dynamics of the DSN is described by a digital spike
map (Dmap, [4] [5]). The Dmap is defined on a set of
points and can be regarded as a digital version of anal-
og one-dimensional map such as the logistic map [6].
Since the domain of the Dmap is a set of finite number
of points, the Dmap cannot generate chaos but a variety
of periodic/transient phenomena. The DSN and Dmap are
well suited for computer-aided precise analysis and FPGA-
based hardware implementation [1].

The Dmap is related to several digital dynamical systems
such as logical/sequential circuits [8], cellular automata [9]
[10], and dynamic binary neural networks [11] [12]. Such
systems are applicable to information compression, signal
processing, and control of switching power converters. Sta-
bility analysis of the Dmap and DSN can contribute not
only to basic study of nonlinear dynamics but also to engi-
eering applications.

2. Digital Spiking Neuron

Fig. 1 (a) illustrates the DSN consisting of two shift reg-
isters connected by a wiring. The left and right shift regis-
ters are referred to as P-cells and X-cells. The P-cells con-
sist of $M$ elements and operates as a pacemaker with period
$M$. Let $P(\tau) \equiv (P_1(\tau), \cdots, P_M(\tau))$ denote the P-cells and
let $P_i(\tau) \in \{0, 1\}$ be the $i$-th element. The dynamics is de-
scribed by

$$P_i(\tau) = \begin{cases} 1 & \text{if } \tau = i + nM \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $i \in \{1, \cdots, M\}$ and $n$ denotes integers.

The X-cells consist of $N$ elements and are state variables
at time. Let $X(\tau) \equiv (X_1(\tau), \cdots, X_N(\tau))$ denote the X-cells and let
$X_j(\tau) \in \{0, 1\}$ be the $j$-th element. An element of the P-
cells is connected to either element of the X-cells in one-
way. The connection is defined by the wiring vector

$$a = (a_1, \cdots, a_M), \quad a_i = j \text{ if } P_i \text{ is connected to } X_j$$

For example, the wiring vector for Fig. 1 (a) is

$$a = (4, 6, 10, 6, 12, 7, 12)$$

Using the wiring vector, we define the base signal with pe-
riod $M$

$$B(\tau) \equiv (B_1(\tau), \cdots, B_N(\tau)), \quad B(\tau + M) = B(\tau)$$

$$B_i(\tau) = \begin{cases} 1 & \text{if } a_i = i \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $\tau \in \{1, 2, \cdots, M\}$ and $i \in \{1, 2, \cdots, N\}$. An example
of $B(\tau)$ is illustrated in Fig. 1 (b). In the X-cells, an initial
condition is assumed to be

$$X_j(1) = \begin{cases} 1 & \text{for some } k \\ 0 & \text{for } j \neq k \end{cases} \quad (3)$$

where $k \in \{1, 2, \cdots, N\}$ and $j \in \{1, 2, \cdots, N\}$. Only one
element can be 1. The dynamics is described by

$$X_{j+1}(\tau + 1) = \begin{cases} 1 & \text{if } X_j(\tau) = 1 \text{ for } j = 1 \sim N \\ 1 & \text{if } X_j(\tau) \neq 1 \text{ and } B_{j+1}(\tau) = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$
If the top element of X-cells is active then the DSN outputs a spike $Y(\tau) = 1$ and generates a spike-train as shown in Fig. 1 (b)

$$Y(\tau + 1) = \begin{cases} 1 & \text{if } X_N(\tau) = 1 \\ 0 & \text{otherwise} \end{cases}$$

(5)

For simplicity, we assume

$$N = 2M - 1, \quad 1 < a_i - 1 \leq M$$

(6)

where $i \in \{1, 2, \cdots, M\}$. In this case, a spike appears once per one clock period and the $n$-th spike appears in the $n$-th interval $[(n - 1)M, \cdots, nM]$ [2] [3]. Let $\theta_n$ denote the $n$-th spike-phase such that

$$Y(\tau + 1) = \begin{cases} 1 & \text{for } \tau = (n - 1)M + \theta_n \\ 0 & \text{otherwise} \end{cases}$$

(7)

where $\theta_n \in \{1, 2, \cdots, M\}$ and $n$ denotes integers. A spike-train is represented by a sequence of spike phases $\{\theta_i\}$. Since the $n$-th spike determines the $(n+1)$-th spike, we can define the digital spike map (Dmap) of the spike-phase:

$$\theta_{n+1} = F(\theta_n), \quad \theta_n \in \{1, 2, \cdots, M\}$$

(8)

An example of the Dmap is shown in Fig. 1 (b). The derivation process and theoretical formula of Dmaps can be found in [1]-[3]. That is, the dynamics of the DSN is integrated into the Dmap. As stated earlier, the Dmap can be regarded as a digital version of analog return maps represented by the logistic map [6]. Although the analog map can generate chaos, the Dmap cannot generate chaos because $M$ is a finite number. However, the Dmap can generate a variety of periodic/transient phenomena as suggested in [4] [5].

3. Super-stabilizing wiring method

In order to consider stabilization of PST, we give several definitions for the Dmap.

**Definition 1:** A point $p \in L_M$ is said to be a periodic point with period $k$ if $p = f^k(p)$ and $f(p)$ to $f^k(p)$ are all different where $f^k$ is the $k$-fold composition of $f$. A sequence of the periodic points $\{p, f(p), \cdots, f^{k-1}(p)\}$ is said to be a periodic orbit (PEO) with period $k$.

A PEO with period $k$ $(f(p) = f^k(p))$ corresponds to a PST with period $MK (Y(\tau + MK) = Y(\tau))$. For example, the PEO with period 4 in Fig. 1 (a) to the PST with period 4 in Fig. 1 (b). Since a PEO in the Dmap is equivalent to a PST in the DSN, we consider stabilization for the PEO in the Dmap instead of the PST for simplicity. For convenience to give definition of stability, we assume that the a period of PEO is at most $M/2$.

**Definition 2:** A point $q \in L_M$ is said to be an eventually periodic point (EPP) with step $k$ if the $q$ is not a periodic point but falls into some periodic point $p$ after $k$ steps: $f^k(q) = p$. An EPP with step 1 is referred to as a direct eventually periodic point (DEPP): $f(q) = p$. An EPP corresponds to an initial spike-position of a transient spike-train to the PST.

**Definition 3:** A PEO is said to be stable if at least one EPP falls into the PEO. A PEO is said to be super-stable if all the EPPs are DEPP falling into the PEO. For example, in Fig. 2, the Dmap has PEO with period 4 and the other 16 - 4 blue points are DEPPs falling into the PEO hence the PEO is super-stable.

Here we introduce the super-stabilizing wiring method (SSWM [3]) for a PEO. As a precondition for the SSWM, we assume that the PEO is given by some algorithm to satisfy some desired characteristics. For example, Ref. [3] has presented a simple evolutionary algorithm that gives a PEO of low autocorrelation.

For simplicity, we explain the SSWM for an example: a PEO with period 3 for $M = 8$ (PST with period 24) in Fig. 1 (a).

$$\alpha_p = \{6, 6, 12\}, \quad \alpha = \{4, 6, 10, 6, 12, 7, 12\}$$

(9)

If other 5 elements are given by the following, we obtain the DSN in Fig. 2 (a).

$$\alpha = \{6 - 1, 6, 6 - 1, 6, 6 + 1, 12, 12 + 1, 12 + 2\}$$

$$= \{5, 6, 5, 6, 7, 12, 13, 14\}$$

(10)

The difference between $\alpha$ and $\alpha$ is rewiring of five blue branches in Fig. 2 (a). In the DSN, the PEO with period 3 is super-stabilized as shown in Fig. 2 (c) where 5 blue points are DEPPs falling directly into the PEO.

In general, if a PEO is given, the PEO can be super-stabilized as the following. First, let the PST with period $pM$ be given by $p$ elements in a wiring vector $\alpha$

$$\alpha_p = \{\alpha_p, \cdots, \alpha_p\} \subset \{\alpha_1, \cdots, \alpha_p, \cdots, \alpha_p\}$$

$$\alpha = \{\alpha_1, \cdots, \alpha_p, \cdots, \alpha_p\}$$

(11)

The following wiring can super-stabilize the PEO.

$$\alpha = \{a_1, \cdots, a_M\}$$

$$a_i = \begin{cases} \alpha_i & \text{if } \alpha_i \in \alpha_p \\ \alpha_j + (i - j) & \text{if } \alpha_i \notin \alpha_p, \quad \alpha_j \in \alpha_p \end{cases}$$

where $j$ is selected arbitrary from $\{p_1, \cdots, p_p\}$.

4. Conclusions

Super-stabilization of periodic spike-trains in the DSN is studied in this paper. In order to visualize the dynamics of DSN, the Dmap is introduced. In order to super-stabilize a desired periodic spike-train, the SSWM is presented.

Future problems include development of the SSWM into various digital dynamical systems, design of basic hardware that can generate super-stable spike-trains, and engineering application of the SSWM.
Figure 1: Digital spiking neuron (DSN). (a) Configuration. (b) Time-domain waveform of the state variable $X(\tau)$ and base signal $B(\tau)$. (c) Digital spike map. Red points construct a PEO and black points are EPPs to the PEO.

Figure 2: DSN after the SSWM. (a) Configuration. (b) Time-domain waveform of state variable and base signal. (c) Digital spike map. Red points construct a PEO with period 3 and blue points are DEPPs to the PEO.
References


Takashi Matsubara† and Kuniaki Uehara†

†Graduate School of System informatics, Kobe University
1-1 Rokko-dai, Nada, Kobe, Hyogo 657–8501, Japan.
Email: matsubara@phoenix.kobe-u.ac.jp, uehara@kobe-u.ac.jp

Abstract—Artificial neural networks with stochastic state transitions, such as Boltzmann machines, have excelled other machine learning approaches in various benchmark tasks. They however require implementation of nonlinear continuous functions and generation of numerous pseudo random numbers, resulting in increase in computational resources. This study proposes a novel implementation method of Boltzmann machine using asynchronous network of cellular automaton-based neurons. The proposed approach requires much less computational resources than traditional implementation approaches since it does not require both the nonlinear continuous functions.

1. Introduction

Artificial neural networks having deep architecture recently have achieved state-of-the-art results in various benchmark tasks (see [1, 2, 3] for review). Some of these successes depend on Boltzmann machines [4, 5]. They are artificial neural networks consisting of bidirectionally connected units with stochastic transitions; they can approximate a given probability distribution by appropriate learning algorithm [6]. As the previous studies mentioned, it requires repeated Gibbs sampling to learn and obtain a distribution [5]. In practice, the Boltzmann machines also require generation of numerous pseudo random numbers for Gibbs sampling and computation of nonlinear probability functions. The architecture of the cutting-edge artificial neural network becomes larger and larger [7] and becomes requiring a dedicated hardware [8, 9].

On the other hand, more biologically plausible neural network model, spiking neural network, also attracts attention as an alternative artificial neural network [10, 11]. Some studies modified spiking neural network models to act as Boltzmann machines [12, 13, 14]. They employed stochastic state transitions or strong noise induction to implement the stochastic units, and lack a perspective of computational efficiency. These studies however imply that a spiking neuron has a potential to act as a stochastic unit.

Recently, an alternative modeling and implementation approach for spiking neural network has been investigated; the nonlinear dynamics of a neuron is modeled as an asynchronous cellular automaton and is implemented as an asynchronous sequential logic circuit [15, 16, 17, 11]. These models have achieved better results in task of reproducing dynamics of mammalian nervous system and required less computational resources than traditional artificial neural networks. Following these previous studies, this paper proposes a type of the hardware-oriented spiking neural networks. Empirical evaluation demonstrates that the proposed model acts as Boltzmann machine and approximates a give probability distribution well despite that it requires less computational resources.

2. Asynchronous Network of Cellular Automaton-based Neuron

This study introduces a type of cellular automaton-based neuron models (ab. CANs) [15, 16, 17, 11]. The dynamic of the CAN in this paper is similar to those of the previously proposed versions but not the same. A CAN is denoted by an index k. The CAN k has a state V_k, which is restricted to the range of [0, 1]. The state V_k can be regarded as a membrane potential from a neuron model viewpoint. The CAN k accepts a periodic internal clock C_k(t) expressed as

\[ C_k(t) = \begin{cases} 1 \text{ if } (t - \theta_k) \pmod{1/f^C_k} = 0, \\ 0 \text{ otherwise,} \end{cases} \]

where \( f^C_k \) is the internal clock frequency, \( \theta_k \) is its initial phase, and \( t \in [0, \infty) \) is the continuous time. The CAN also accepts multiple external binary inputs \( S_i(t) \in [0, 1] \) They can be regarded as pre-synaptic action potentials from a neuron model viewpoint. The accepted inputs generate the following signal \( U_k \):

\[ U_k(t) = \sum_i G_{kj} S_i(t). \] (1)

where \( G_{kj} \) can be regarded as a synaptic weight from the pre-synaptic action potential \( S_i(t) \) to the CAN k, and \( G_{kj} > 0 \) (\( G_{kj} < 0 \)) implies excitation (inhibition). At the rising edge of the internal clock \( C_k(t) \), the membrane potential \( V_k \) is updated as

\[ V_k(t^+) = \begin{cases} V_k(t) - c_k + U_k(t) + \xi(t) \text{ if } C_k(t)=1, \\ V_k(t) \text{ otherwise,} \end{cases} \]
where the variable $t^*$ denotes the moment just after $t$, i.e.,
$t^* = \lim_{\epsilon \to 0} t + \epsilon$, the parameter $c_k$ represents a leak current, $\xi(t)$ has a noise term. When the membrane potential $V_k$ reaches or exceeds 1.0, the membrane potential $V_k$ is immediately reset and the CAN $k$ generates an output $Y_k(t) = 1$ as

$$V_k(t^*) = \begin{cases} V_k(t) \pmod{1} & \text{if } V_k(t) \geq 1, \\ V_k(t) & \text{otherwise}, \end{cases}$$

and

$$Y_k(t^*) = \begin{cases} 1 & \text{if } V_k(t) \geq 1, \\ 0 & \text{if } C(t) = 1, \\ Y_k(t) & \text{otherwise}. \end{cases}$$

An asynchronous recurrent networks of CANs (ab. ANCAN) [11] consists of $n$ CANs. A CAN $l$ is connected to another CAN $k$ via a synaptic weight $G_{lk}$. An action potential $Y_l(t) = 1$ generated by the CAN $l$ is delivered to the CAN $n_0^l$ and is accepted as a pre-synaptic action potential $S_l(t) = 1$, i.e.,

$$S_l(t) = Y_l(t).$$

### 2.1. Conversion from Boltzmann Machine

The detailed description of dynamics of the Boltzmann machine is outside of scope of this paper and therefore is omitted. This paper focuses on a Bernoulli restricted Boltzmann machine, consisting of $n_v$ visible units and $n_h$ hidden units; they have bias terms $b_0$ and $b_h$ and are connected via synaptic weights $W_{ij}$. Each unit has a binary state; 0 or 1. The target ANCAN was constructed with $n_v$ CANs corresponding to the visible units and $n_h$ CANs corresponding to the hidden units. The leak current $c_k$ corresponding to a unit $k$ is set to $-b_k$. The synaptic weights $G_{ij}$ are set to the corresponding synaptic weights $W_{ij}$. The internal clock frequency $f^C_k$ is randomly chosen from a uniform distribution $\mathcal{U}(1,2)$. For comparison, the synchronous version of the ANCAN, called SANCAN, is also prepared; internal clock frequency is uniformly 1 and the initial phase $\theta_k$ of a visible unit is 0 and that of a hidden unit is 0.5. For mimicking sigmoid function $\sigma(u) = \left(1 + \exp(-u)\right)^{-1}$, the noise term $\xi$ is set to follow a normal distribution $\mathcal{N}(\frac{\xi}{3}, 3)$.

### 3. Results

#### 3.1. Activation Function

When a CAN accepts the fixed inputs $U_k(t) = x$ and the probability of action potential is denoted as $y = P(Y = 1)$, the empirical relationship between $x$ and $y$ is depicted in Fig. 1 (a). For comparison, the sigmoid function $y = \left(1 - \exp(-x)\right)^{-1}$ and its piecewise linear approximation called

![Figure 1: (a) The CAN and the other activation functions.](image)

- 635 -
in several ways. The Boltzmann machine was implemented on a general purpose computer with 64-bit floating point numbers, the sigmoid function, and the Mersenne twister pseudo-random number generator and was used for the original Boltzmann machine. The other Boltzmann machines were implemented with fixed-point numbers of scaling factor $2^k$. The activation functions were the sigmoid function, the PLAN, and the proposed CAN. For the sigmoid function and the PLAN, the M-sequence random number generator (ab. M-seq. RNG) of 16-bit was used to generate uniform distribution $\mathcal{U}(0, 1)$. For the proposed CAN, the M-seq. RNG of 16-bit was also used to generate the noise term $\xi$ following binomial distribution $\mathcal{B}(12, 0.5)$, where $\hat{\xi} + \frac{3}{2} - 6$ is a good approximation of the noise term $\xi$ following a normal distribution $\mathcal{N}(\frac{3}{2}, 3)$. The M-seq. RNG was updated according to the internal clock with the frequency of 1.

After the Boltzmann machines were initialized and the neurons were updated repeatedly, the Boltzmann machines reached a stationary distribution. In this paper, the outputs of the first five visible neurons and their distribution were focused. The Kullback-Leibler divergence $D_{KL}$ was used to measure the similarity between the stationary distributions of the original Boltzmann machine and the implemented Boltzmann machine. The smaller Kullback-Leibler divergence $D_{KL}$ implies the better accuracy of the approximation. The number of neurons was set to $n = 5, 20, 100$. When $n \leq 20$, the stationary distribution of the original Boltzmann machine was theoretically obtained. Otherwise the empirical distribution of $10^7$ samples from the original Boltzmann machine was used. The average Kullback-Leibler divergences $D_{KL}$ obtained from 10 trials is shown in Fig. 2 and is summarized in Table 1. When $n = 5$, the sigmoid function demonstrated the best performance. The PLAN and the ANCAN were comparable, while the SNCAN had a worse performance. When $n = 20$ and $n = 100$, the ACAN got performance comparable to the sigmoid function and excelled the PLAN. Remarkably, the Kullback-Leibler divergence $D_{KL}$ of the ANCAN kept decreasing after sampling $10^6$ times, while that of the PLAN converged before sampling $10^5$ times.

### 3.3. Implementation

The Boltzmann machines were also implemented on an field programmable gate array (FPGA) device with the fixed-point numbers with scaling factor $2^k$. Xilinx FPGA Kintex-7 XC7K325T-2FFG900C mounted on the Kintex-7 FPGA KC705 Evaluation Kit [19] was used. A bitstream file for the FPGA configuration was generated by the Xilinx design software environment ISE 14.7. The corresponding implementation cost, i.e., the number of the occupied slices on the FPGA devices, are also shown in Table 1. Straight-forward implementation of the sigmoid function is trouble some and thus is omitted. When $n = 5$, the ANCAN and the SNCAN require computational resources less than half that of the PLAN. When $n = 20$, they reduced the implementation cost by 30 %.

### 4. Discussion

These results suggest that the ANCAN is a better approximation and requires much less computational resources when compared to the Boltzmann machines with the sigmoid function or the PLAN function. This study
Table 1: Comparison between the Boltzmann Machines.

<table>
<thead>
<tr>
<th>function</th>
<th>bit length</th>
<th>RNG</th>
<th>$D_{KL}$</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>sigmoid</td>
<td>64-bit float</td>
<td>Mersenne twister</td>
<td>$&lt; 1.68 \times 10^{-5}$</td>
<td>$2.00 \times 10^{-5}$</td>
</tr>
<tr>
<td>sigmoid</td>
<td>8-bit fixed</td>
<td>M-seq. RNG of 16bit</td>
<td>$1.28 \times 10^{-4}$</td>
<td>$2.15 \times 10^{-4}$</td>
</tr>
<tr>
<td>PLAN</td>
<td>8-bit fixed</td>
<td>M-seq. RNG of 16bit</td>
<td>$3.06 \times 10^{-4}$</td>
<td>$5.43 \times 10^{-4}$</td>
</tr>
<tr>
<td>CAN (sync.)</td>
<td>8-bit fixed</td>
<td>M-seq. RNG of 16bit</td>
<td>$1.67 \times 10^{-3}$</td>
<td>$1.10 \times 10^{-3}$</td>
</tr>
<tr>
<td>CAN (async.)</td>
<td>8-bit fixed</td>
<td>M-seq. RNG of 16bit</td>
<td>$3.41 \times 10^{-3}$</td>
<td>$2.62 \times 10^{-4}$</td>
</tr>
</tbody>
</table>

was partially supported by the KAKENHI (16K12487 and 26280040), Kayamori Foundation of Information Science Advancement, and The Nakajima Foundation.

References


Development of Discrete Mechanics for 2-dimensional Distributed Parameter Mechanical Systems and Its Application to Vibration Suppression Control of a Film

Tatsuya Kai and Kouhei Yamaki

†Graduate School of Industrial Science and Technology, Tokyo University of Science
6-3-1 Niijuku, Katsushika-ku, Tokyo 125-8585, Japan

Abstract—This paper develops a new discretizing method for 2-dimensional distributed parameter mechanical systems, called “discrete mechanics”, and considers its applications to control theory. Especially, a new control method based on discrete mechanics and nonlinear optimization is proposed. The new method is applied to the vibration suppression problem of a film as a physical example. From numerical simulation results, it turns out that vibration of the film is suppressed by control inputs and the whole of the film is stabilized.

1. Introduction

Control of distributed parameter systems is well-known to be one of the most challenging problems in control theory. In general, control of distributed parameter systems is more difficult than that of concentrated constant systems, because distributed parameter systems are represented by infinite-dimensional equations and a lot of actuators and sensors are needed for control. There are generally two kinds of ways to control distributed parameter systems; analytic methods and numerical methods. Especially, numerical methods are powerful tools and a lot of work have been done so far. The authors have been developing a new discretizing method of distributed parameter mechanical systems called “discrete mechanics,” which is an extension of the case for concentrated constant systems [1, 2, 3, 4, 5]. The main concept of discrete mechanics is that we first discretize some fundamental concepts of classical physics such as Lagrangian and Hamilton’s principal, and then derive discrete euler-Lagrange equations by using discrete Hamilton’s principle. In [6, 7], for the 1-dimensional case, a new control method based on discrete mechanics and nonlinear optimization has been proposed and its application potentiality has been confirmed by numerical simulations.

This study aims at development of discrete mechanics for 2-dimensional distributed parameter mechanical systems as an extension of the previous work for the 1-dimensional case [6, 7], and an application to vibration suppression control of a film. First, Section 2 describes details on discrete mechanics for 2-dimensional distributed parameter mechanical systems. Next, Section 3 presents a new control method via a blending method of discrete mechanics and nonlinear optimization. Then, in Section 4, the vibration suppression control problem of a film is considered as a physical example, and some numerical simulations are shown in order to check the effectiveness of the new method.

2. Discrete Mechanics for 2-dimensional Distributed Parameter Mechanical Systems

This section derives some important concepts on discrete mechanics for 2-dimensional distributed parameter mechanical systems. Let us denote the time variable as \( t \in \mathbb{R} \) and the position of the 2-dimensional space as \((x, y) \in \mathbb{R}^2\). We also refer a displacement of the system at the time \( t \) and the position \((x, y)\) as \( u(t, x, y) \in \mathbb{R} \), and \( u(t, x, y) \) with a subscript indicates partial derivative of \( u(t, x, y) \) with respect to the subscript, e.g. \( u_t, u_x, u_y \). In this paper, we deal with a continuous Lagrangian density which includes through first-order partial derivative of \( u(t, x, y) \) as

\[
L'(t, x, u, u_t, u_x, u_y).
\]

Next, we consider discretization of variables. As shown in Fig. 1, the time variable \( t \) and the position \((x, y)\) are discretized with sampling intervals \( h, d_x, \) and \( d_y \) as

\[
t \approx kh \quad (k = 1, 2, \ldots, K - 1, K),
x \approx dx_l \quad (l = 1, 2, \ldots, L - 1, L),
y \approx dy_m \quad (l = 1, 2, \ldots, M - 1, M),
\]

where \( k, l, \) and \( m \) are indices of \( t, x, \) and \( y \), respectively.

Now, we use a new notation \( U_{k,l,m} \in \mathbb{R} \) as a discrete version of the displacement of the system at the time step \( k \) and the position \((l, m)\). Then, we assume that the continuous displacement of the system at the time \( t \) and the position \((x, y)\): \( u(t, x, y) \) is represented as

\[
u(t, x, y) \approx (1 - \alpha)(1 - \beta_1)(1 - \beta_2)U_{k,l,m} + (1 - \alpha)(1 - \beta_1)\beta_2 U_{k,l,m+1} + (1 - \alpha)\beta_1(1 - \beta_2)U_{k+1,l,m} + (1 - \alpha)\beta_1\beta_2 U_{k+1,l,m+1} + \alpha(1 - \beta_1)(1 - \beta_2)U_{k+1,l,m+1} + \alpha(1 - \beta_1)\beta_2 U_{k+1,l,m+1} + \alpha\beta_1(1 - \beta_2)U_{k+1,l,m+1} + \alpha\beta_1\beta_2 U_{k+1,l,m+1},
\]

with 8 displacement variables: \( U_{k,l,m}, U_{k,l,m+1}, U_{k,l+1,m}, U_{k+1,l,m}, U_{k+1,l,m+1}, U_{k,l+1,m+1}, U_{k+1,l+1,m}, U_{k+1,l+1,m+1} \).
where \( \alpha, \beta_1, \beta_2 \in \mathbb{R} \) are dividing parameters \((0 < \alpha, \beta_1, \beta_2 < 1)\). Partial derivatives of \( u(t, x, y) \) are also represented by

\[
\begin{align*}
  u_i(t, x, y) &\approx \frac{U_{k+1,m} - U_{k,m}}{h}, \\
  u_i(t, x, y) &\approx \frac{U_{k+1,m} - U_{k,m}}{d_i}, \quad \text{(4)} \\
  u_i(t, x, y) &\approx \frac{U_{k,m+1} - U_{k,m}}{d_y}.
\end{align*}
\]

By substituting (2)–(4) into (1) and multiplying it by \(hd_id_y\), we define “a discrete Lagrangian density” as

\[
L^d_{k,l,m} \approx hd_id_y L^e. \quad \text{(5)}
\]

We also define “a discrete action sum” as

\[
S^d(U) := \sum_{k=1}^{K-1} \sum_{l=1}^{L-1} \sum_{m=2}^{M-1} L^d_{k,l,m} \quad \text{(6)}
\]

and consider “a discrete variation” as

\[
\delta S^d(U) := S^d(U + \delta U) - S^d(U),
\]

where \( \delta U \) is a variation of \( U \) and satisfies the boundary conditions:

\[
\begin{align*}
  \delta U_{1,m} &= \delta U_{K,M} = \delta U_{K,l,m} = \delta U_{k,l,1} = \delta U_{k,M} = 0. \\
  &\quad \text{(k = 1, \cdots, K; l = 1 \cdots, L; m = 1 \cdots, M)}
\end{align*}
\]

As an analogy of Hamilton’s principle in the continuous version, we consider “discrete Hamilton’s principle” and it states that “only a motion such that the discrete action sum (6) is stationary, that is, \( S^d(U) = 0 \), can be realized.” By applying discrete Hamilton’s principle to the discrete action sum (6) and calculating in details, we can derive “discrete Euler-Lagrange equations” as the following (due to limitations of space, the proof is omitted).

**Theorem 1**: For the discrete Lagrangian density \( L^d_{k,l,m} \) (5) for 2-dimensional distributed parameter mechanical systems, the discrete Euler-Lagrange equation that satisfies discrete Hamilton’s principle is given by

\[
\begin{align*}
  \frac{\partial L^d_{k-1,l-1,m-1}}{\partial U_{k,l,m}} + &\; \frac{\partial L^d_{k-1,l-1,m}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k-1,l,m-1}}{\partial U_{k,l,m}} + \\
  + &\; \frac{\partial L^d_{k,l-1,m-1}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k,l-1,m}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k,l,m-1}}{\partial U_{k,l,m}} + \\
  + &\; \frac{\partial L^d_{k+1,l,m-1}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k+1,l,m}}{\partial U_{k,l,m}} = 0
\end{align*}
\]

\((k = 2, \cdots, K - 1; l = 2 \cdots, L - 1; m = 2 \cdots, M - 1)\)

We can calculate all the KLM displacements \( U_{k,l,m} \) by using the discrete Euler-Lagrange equations (9) under suitable initial and boundary conditions. In addition, the discrete Euler-Lagrange equations (9) are generally nonlinear and implicit, and hence we need some numerical solutions for nonlinear equations such as Newton’s method in order to calculate all the displacements of the system.

### 3. Optimal Control Method via Discrete Mechanics

In this section, a nonlinear control problem for a mathematical model derived by discrete mechanics is formulated, and a solution method of the problem is considered. First, the setting on control inputs is shown. Denote a control input at the time step \( k \) and the position \((l, m)\) as \( F_{k,l,m} \in \mathbb{R} \). If an actuator is not installed at the position \((l, m)\), we set \( F_{k,l,m} = 0 \) \((k = 1, \cdots, K)\). We also denote a set of indices \((l, m)\) such that actuators are installed as \( \Delta \). Thus, the discrete Euler-Lagrange equations with control inputs are given by

\[
\begin{align*}
  \frac{\partial L^d_{k-1,l-1,m-1}}{\partial U_{k,l,m}} + &\; \frac{\partial L^d_{k-1,l-1,m}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k-1,l,m-1}}{\partial U_{k,l,m}} + \\
  + &\; \frac{\partial L^d_{k,l-1,m-1}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k,l-1,m}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k,l,m-1}}{\partial U_{k,l,m}} + \\
  + &\; \frac{\partial L^d_{k+1,l,m-1}}{\partial U_{k,l,m}} + \frac{\partial L^d_{k+1,l,m}}{\partial U_{k,l,m}} = F_{k,l,m}
\end{align*}
\]

\((k = 2, \cdots, K - 1; l = 2 \cdots, L - 1; m = 2 \cdots, M - 1)\)

In this study, the next control problem is dealt with for the discrete Euler-Lagrange equations with control inputs (10).

**Problem 1**: For the discrete Lagrangian density \( L^d_{k,l,m} \) (5) and the discrete Euler-Lagrange equation with control inputs (10), find control inputs \( F_{k,l,m} \) \((k = 2, \cdots, K - 1; (l, m) \in \Delta)\) that make all the specified displacements \( U_{k,l,m} \) \((k = K, \cdots, K; l = 1, \cdots, L, m = 1, \cdots, M)\) converge to 0. \( \square \)

In order to solve Problem 1, we consider an optimal control approach. Using weight parameters \( a, b, c \), we set an
evaluation function as

\[ J(U, F) = a \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{m=1}^{M} U_{k,l,m}^2 + b \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{m=1}^{M} U_{k,l,m}^2 + c \sum_{k=1}^{K-2} \sum_{l=1}^{L} \sum_{m=1}^{M} F_{k,l,m}^2 \]  

(11)

where the first and second terms evaluate the displacements from \( k = 1 \) to \( k = \kappa - 1 \) and ones from \( k = \kappa \) to \( k = K \), respectively, and the third term evaluates the values of control inputs. It can be expect that we can make all the specified displacements converge to 0. by minimizing the evaluation function (11). The optimal control problem for the discrete Euler-Lagrange equation with control inputs (10) can be formulated as

\[
\min_{U,F} \ (11), \\
\text{subject to} \ (10), \\
given \text{ initial conditions, boundary conditions.}
\]  

(12)

The optimal control problem (12) can be referred as a finite-dimensional nonlinear optimization problem with constraints, and hence we can solve it by numerical solutions such as “the sequential quadratic programming method” [8]. It is known that the sequential quadratic programming method can be applied to a relatively large-scale problems and effectively obtain an optimal or near-optimal solution.

4. Vibration Suppression Control of Film

This section deals with an application to a physical system “a film,” and confirms the effectiveness of the proposed control method via numerical simulations. It is assumed that the shape of the film is rectangle and the film is clamped at four sides as illustrated in Fig. 2. Denote the 2-dimensional position of the film as \((x, y)\) and the displacement of the film at time \(t\) and the position \((x, y)\) as \(u(t, x, y)\). Physical parameters of the film are set as \(\rho\): a density of the film, \(E\): tension of the film. Then, the continuous Lagrangian density of the film is given by

\[ L^c = \frac{1}{2} \rho u_t^2 - \frac{1}{2} E (u_x^2 + u_y^2). \]

(13)

Note that the continuous Lagrangian density (13) contains through first-order partial derivative \(u_t, u_x, u_y\).

Discretization setting is the same as the one explained in the previous section. From (13), we have the discrete Lagrangian density of the film as

\[
L^d_{k,l,m} = \frac{h d_x d_y}{2} \left\{ \rho \left( \frac{U_{k+1,l,m} - U_{k,l,m}}{h} \right)^2 - E \left( \frac{U_{k+1,l,m} - U_{k,l,m}}{d_x} \right)^2 \right\} 
- E \left( \frac{U_{k,l+1,m} - U_{k,l,m}}{d_y} \right)^2 - E \left( \frac{U_{k,l,m+1} - U_{k,l,m}}{d_y} \right)^2, \]

(14)

and hence from (10) we obtain the discrete Euler-Lagrange equation of the film as

\[
- \frac{\rho}{h^2} (U_{k+1,l,m} + U_{k-1,l,m}) + \frac{E^2}{d_x^2} (U_{k,l+1,m} + U_{k,l-1,m}) + c \frac{E^2}{d_y^2} (U_{k,l,m+1} + U_{k,l,m-1}) + 2 \left( \frac{\rho}{h^2} - \frac{E}{d_x^2} - \frac{E}{d_y^2} \right) U_{k,l,m} = F_{k,l,m}. \]

(15)

We see that (15) contains 7 displacement variables \( U_{k-1,l,m}, U_{k+1,l,m}, U_{k,l+1,m}, U_{k,l,m-1}, U_{k,l,m+1}, U_{k,l+1,m}, U_{k+1,l,m} \). In computation of numerical solutions, a numerical stability condition called “a von Neumann condition” is quite important [9]. The next proposition gives a von Neumann condition for the discrete Euler-Lagrange equation of the film. (due to limitations of space, the proof is omitted.)

Proposition 1: A von Neumann condition such that the discrete Euler-Lagrange equation of the film (15) is numerically stable is given by

\[ 0 < \frac{E}{\rho} \frac{1}{h^2} \left( \frac{1}{d_x^2} + \frac{1}{d_y^2} \right) \leq 1. \]

(16)

Then, a numerical simulation is performed by the proposed control method. We assume that the number of actuators is 4 and they are installed at four corners of the film as illustrated in Fig. 2. The parameters are set as the physical parameters: \(\rho = 1, E = 1\), the sampling intervals: \(h = 0.01, d_x = 0.1, d_y = 0.1\), the total steps: \(K = 50, L = 20, M = 20\), the set of actuator indices: \(A = \{2, 2\}, \{2, 19\}, \{19, 2\}, \{19, 19\}\), the start time step of stabilization: \(\kappa = 45\), the weight parameters of evaluation function: \(a = 1, b = 3000, c = 1\). Note that these parameters satisfy the von Neumann condition (16).

Fig. 3 shows the simulation result on a 3D plot of the displacements of the film \( U_{k,l,m} \). From this figure, it can be confirmed that all the displacements of the film converge to 0, and hence vibration suppression control is achieved. It is also possible to stabilize the film at earlier time step by tuning the parameters in the evaluation function: \(a, b, c\).
5. Conclusions

This study has developed discrete mechanics for 2-dimensional distributed parameter mechanical systems and a new control method by blending of discrete mechanics and nonlinear optimization. A numerical simulation for a film has shown that vibration of the film is suppressed by control inputs, and then the whole of the film is stabilized by the proposed method.

The future work includes the next topics; theoretical analysis on discrete Euler-Lagrange equations and development of feedback-type controllers.

References


Figure 3: Snapshot of Film
A Hardware-Efficient Gene Network Model
based on Asynchronous Bifurcation Processor

Takuya Yoshimoto and Hiroyuki Torikai†
†Department of Computer Science, Kyoto Sangyo University
Motoyama, Kamigamo, Kita-Ku, Kyoto 603-8555, Japan
Email: g1345521@cse.kyoto-su.ac.jp, torikai@cse.kyoto-su.ac.jp

Abstract—In this paper, a novel gene network model based on an asynchronous bifurcation processor is presented. It is shown that an isolated gene network model generates periodic oscillations of mRNA, protein, inhibitor, and neuropeptide concentrations. Then a network of the gene network (corresponding to a network of cells) is presented, where each gene network is coupled via a mean field of the neuropeptides. It is shown that the network of the cells can reproduce typical nonlinear phenomena related to circadian rhythms such as mutual synchronization of coupled gene networks and forced synchronization of coupled gene networks to a periodic light stimulation.

1. Introduction

The circadian rhythm is a biological rhythmic phenomena, where mammalian circadian rhythms typically have periods of about 24-h (24 hours). The circadian rhythm is generated by a network of suprachiasmatic nucleus (SCN) neurons [1]-[3]. A gene network in the single SCN neuron forms a closed-loop nonlinear dynamical system, which leads to periodic oscillations of mRNA, protein, and neurotransmitter concentrations. If the SCN neurons are isolated, oscillation periods of the mRNA concentrations have a relatively wide distribution. However, in the suprachiasmatic nucleus, the SCN neurons are coupled via neurotransmitters and the oscillation periods of the mRNA concentrations have a sharp distribution near (but not identical with) 24-h. In addition, if a light-and-dark stimulation with period 24-h is applied to a mammalian visual system, the suprachiasmatic nucleus is affected by the stimulation and then the periods of the mRNA concentrations in the coupled SCN neurons are sharply locked to 24-h. It should be emphasized that the gene networks play important roles not only in the generation of the circadian rhythm but also in many brain functions [3], i.e., investigation of the circadian rhythm is an entrance into studies of roles of gene networks in brain functions. On the other hand, our group has been developing a neural system modeling approach based on the nonlinear dynamics of an asynchronous cellular automaton, where nonlinear dynamics (especially, bifurcations) of neural systems are reproduced by the asynchronous cellular automaton with low hardware cost [4]-[6]. Our group is conceptually referring to such a hardware platform as ”asynchronous bifurcation processor (ABP).” So, in this paper, a novel gene network model based on the ABP is presented. It is shown that the network can reproduce generations of typical circadian rhythms.

2. ODE Gene Network Model

Let

\[ t \in \mathbb{R} \]

represent a continuous time and let

\[ X_i \in \mathbb{R}^+ \], \[ Y_i \in \mathbb{R}^+ \], \[ Z_i \in \mathbb{R}^+ \], \[ V_i \in \mathbb{R}^+ \]

represent mRNA concentration of a clock gene, the resulting protein concentration, concentration of nuclear form of the protein (inhibitor concentration), and neuropeptide concentration, respectively. A basic ODE gene network model is described by the following set of equations [1].

\[
\frac{dX_i}{dt} = v_1 \frac{K_1^p}{K_1^p + Z_i} - v_2 \frac{X_i}{K_2 + X_i} + v_c \frac{KF}{K_c + KF} + L,
\]

\[
\frac{dY_i}{dt} = k_3 X_i - v_4 \frac{Y_i}{K_4 + Y_i},
\]

\[
\frac{dZ_i}{dt} = k_5 Y_i - v_6 \frac{Z_i}{K_6 + Z_i},
\]

\[
\frac{dV_i}{dt} = k_7 X_i - v_8 \frac{V_i}{K_8 + V_i},
\]

\[
F = \frac{1}{D} \sum_{i=1}^{D} V_i,
\]

where \( L \) represents a light stimulation with period 24-h; \( D \) represents the number of cells; \( F \) represents a mean field of the neuropeptide; and \( v_1, \cdots, v_8, K_1, \cdots, K_8, v_c, \) and \( K_c \) are parameters. Fig. 1 shows typical time waveforms of the ODE gene network model. Fig. 1(a) shows time waveforms of concentrations \( X_1 \) and \( X_2 \) of mRNAs in two cells. In Fig. 1(b), the cells are coupled via the mean field \( F \) of the neuropeptide. In this case, the mRNA concentrations \( X_1 \) and \( X_2 \) are synchronized. In Fig. 1(c), the coupled cells accept an external light stimulation \( L \) with period 24-h.
this case, the periods of oscillations of the mRNA concentrations $X_1$ and $X_2$ coincide with the period 24-h of the light stimulation $L$.

3. Gene Network Model based on Asynchronous Bifurcation Processor

In this section, a gene network model based on the asynchronous bifurcation processor (ABP) is proposed. The model has the following discrete states.

$$X_i \in A_N = \{0, \ldots, N - 1\}, \quad Y_i \in A_N, \quad Z_i \in A_N, \quad V_i \in A_N,$$

$$P_i \in A_M = \{0, \ldots, M - 1\}, \quad Q_i \in A_M, \quad R_i \in A_M, \quad S_i \in A_M,$$

where $N$ and $M$ are positive integers, which determine the resolution of the state space. Same as the ODE model, the discrete states $X_i$, $Y_i$, $Z_i$, and $V_i$ correspond to mRNA concentration of a clock gene, the resulting protein concentration, concentration of nuclear form of the protein (inhibitor concentration), and neuropeptide concentration, respectively. The other discrete states $P_i$, $Q_i$, $R_i$, and $S_i$ are used to control velocities of the discrete states $X_i$, $Y_i$, $Z_i$, and $V_i$, respectively. In order to realize a vector field of a gene network, the following discrete functions are introduced.

$$F_{X_i}(X_i, Y_i, Z_i, V_i) = \begin{cases} 
M - 1 & \text{if } \text{Int}(\frac{i}{g_X(X_i, Y_i, Z_i, V_i)}) \geq M - 1, \\
-(M - 1) & \text{if } \text{Int}(\frac{i}{g_X(X_i, Y_i, Z_i, V_i)}) \leq -(M - 1), \\
\text{Int}(\frac{i}{g_X(X_i, Y_i, Z_i, V_i)}) & \text{otherwise}.
\end{cases}$$

$$F_{Y_i}(X_i, Y_i, Z_i, V_i) = \begin{cases} 
M - 1 & \text{if } \text{Int}(\frac{i}{g_Y(X_i, Y_i, Z_i, V_i)}) \geq M - 1, \\
-(M - 1) & \text{if } \text{Int}(\frac{i}{g_Y(X_i, Y_i, Z_i, V_i)}) \leq -(M - 1), \\
\text{Int}(\frac{i}{g_Y(X_i, Y_i, Z_i, V_i)}) & \text{otherwise}.
\end{cases}$$

$$F_{Z_i}(X_i, Y_i, Z_i, V_i) = \begin{cases} 
M - 1 & \text{if } \text{Int}(\frac{i}{g_Z(X_i, Y_i, Z_i, V_i)}) \geq M - 1, \\
-(M - 1) & \text{if } \text{Int}(\frac{i}{g_Z(X_i, Y_i, Z_i, V_i)}) \leq -(M - 1), \\
\text{Int}(\frac{i}{g_Z(X_i, Y_i, Z_i, V_i)}) & \text{otherwise}.
\end{cases}$$

$$F_{V_i}(X_i, Y_i, Z_i, V_i) = \begin{cases} 
M - 1 & \text{if } \text{Int}(\frac{i}{g_V(X_i, Y_i, Z_i, V_i)}) \geq M - 1, \\
-(M - 1) & \text{if } \text{Int}(\frac{i}{g_V(X_i, Y_i, Z_i, V_i)}) \leq -(M - 1), \\
\text{Int}(\frac{i}{g_V(X_i, Y_i, Z_i, V_i)}) & \text{otherwise}.
\end{cases}$$

Figure 1: Typical waveforms of the ODE gene network model [1]. The parameter values are $K = 0.5$, $v_1 = 0.7/0.73$, $K1 = 1.0/1.01$, $n = 4.0$, $v_2 = 0.35/0.33$, $K2 = 1.0/0.98$, $k3 = 0.7/0.72$, $v4 = 0.35$, $K4 = 1.0/1.01$, $k5 = 0.7/0.73$, $v6 = 0.35$, $K6 = 1.0/1.04$, $k7 = 0.35$, $v8 = 1.0/0.97$, $K8 = 1.0/1.01$, $v_c = 0.4/0.41$, and $K_c = 1.0/1.05$. (a) Time waveforms of mRNA concentrations $X_1$ and $X_2$ of two uncoupled cells. (b) Synchronized time waveforms of mRNA concentrations $X_1$ and $X_2$ of two coupled cells. (c) Time waveforms of mRNA concentrations $X_1$ and $X_2$ of two coupled cells under external stimulation of light. The mRNA concentrations $X_1$ and $X_2$ are synchronized to the light.
where
\[
g(X_i, Y_i, Z_i, V_i) = l(v_1 \frac{K^n_i}{K^n_1 + Z^n_i} - v_2 \frac{X_i}{K_2 + X_i} + v_3 \frac{K_F}{K_c + K_F + L}),
g(Y_i, Z_i, V_i) = l(k_3 X_i - v_5 \frac{Y_i}{K_n + Y_i}),
g(X_i, Y_i, Z_i, V_i) = l(k_5 Y_i - v_5 \frac{Z_i}{K_6 + Z_i}),
g(X_i, Y_i, Z_i, V_i) = l(k_7 X_i - v_5 \frac{V_i}{K_S + V_i}),
\]
\[
F = \frac{1}{D} \sum_{i=1}^{D} V_i.
\]
The presented model accepts the following four periodic internal clocks \(C_X(t), C_Y(t), C_Z(t)\) and \(C_V(t)\) with periods \(T_X > 0, T_Y > 0, T_Z > 0,\) and \(T_V > 0\) respectively, which are generated by uncoupled clock generators.
\[
C_X(t) = \begin{cases} 
1 & \text{if } t = 0, 2T_X, 3T_X, \ldots, \\
0 & \text{otherwise}.
\end{cases}
\]
\[
C_Y(t) = \begin{cases} 
1 & \text{if } t = 0, 2T_Y, 3T_Y, \ldots, \\
0 & \text{otherwise}.
\end{cases}
\]
\[
C_Z(t) = \begin{cases} 
1 & \text{if } t = 0, 2T_Z, 3T_Z, \ldots, \\
0 & \text{otherwise}.
\end{cases}
\]
\[
C_V(t) = \begin{cases} 
1 & \text{if } t = 0, 2T_V, 3T_V, \ldots, \\
0 & \text{otherwise}.
\end{cases}
\]
These clocks trigger the following asynchronous transitions of the discrete states.
\[
\text{If } C_X(t) = 1, \text{ then}
\]
\[
P(t) := \begin{cases} 
P(t) + 1 & \text{if } P(t) < |F_X|, \\
0 & \text{if } P(t) \geq |F_X|.
\end{cases}
\]
\[
\text{If } C_Y(t) = 1, \text{ then}
\]
\[
Q(t) := \begin{cases} 
Q(t) + 1 & \text{if } Q(t) < |F_Y|, \\
0 & \text{if } Q(t) \geq |F_Y|.
\end{cases}
\]
\[
\text{If } C_Z(t) = 1, \text{ then}
\]
\[
R(t) := \begin{cases} 
R(t) + 1 & \text{if } R(t) < |F_Z|, \\
0 & \text{if } R(t) \geq |F_Z|.
\end{cases}
\]
\[
\text{If } C_V(t) = 1, \text{ then}
\]
\[
S(t) := \begin{cases} 
S(t) + 1 & \text{if } S(t) < |F_V|, \\
0 & \text{if } S(t) \geq |F_V|.
\end{cases}
\]
\[
\text{If } C_X(t) = 1, \text{ and } P(t) \geq |F_X|, \text{ then}
\]
\[
\text{If } C_Y(t) = 1, \text{ and } Q(t) \geq |F_Y|, \text{ then}
\]
\[
\text{If } C_Z(t) = 1, \text{ and } R(t) \geq |F_Z|, \text{ then}
\]
\[
\text{If } C_V(t) = 1, \text{ and } S(t) \geq |F_V|, \text{ then}
\]
Fig. 2 shows typical time waveforms of the proposed model. Fig. 2(a) shows time waveforms of concentrations \(X_1\) and \(X_2\) of mRNAs in two uncoupled cells. In Fig. 2(b), the cells are coupled via the mean field \(F\) of the neuropeptide. In this case, the mRNA concentrations \(X_1\) and \(X_2\) synchronize. In Fig. 2(c), the coupled cells accept an external light stimulation \(L\) with a period corresponding to 24-h. In this case, the periods of oscillations of the mRNA concentrations \(X_1\) and \(X_2\) coincide with the period of the light stimulation \(L\). Comparing Fig. 2 with Fig. 1, it can be confirmed that the proposed model can reproduce generations of the typical circadian rhythms.

4. Conclusion

In this paper, the gene network model based on the ABP is proposed. It was shown that the model can reproduce the features of the circadian rhythm such as the mutual synchronization phenomena of coupled gene networks and
Figure 2: Typical waveforms of the proposed gene network model. The parameter values are $l = 10000$, $K = 0.8$, $v_1 = 0.7/0.73$, $K_1 = 1.0/1.01$, $n = 4.0$, $v_2 = 0.35/0.33$, $K_2 = 1.0/0.98$, $k_3 = 0.7/0.72$, $v_4 = 0.35$, $K_4 = 1.0/1.01$, $k_5 = 0.7/0.73$, $v_6 = 0.35$, $K_6 = 1.0/1.04$, $k_7 = 0.35$, $v_8 = 1.0/0.97$, $K_8 = 1.0/1.01$, $w_c = 0.4/0.41$, and $Kc = 1.0/1.05$. $T_X = 18$, $T_Y = 17$, $T_Z = 4$, and $T_V = 2$.

(a) Time waveforms of mRNA concentrations $X_1$ and $X_2$ of two uncoupled cells. (b) Synchronized time waveforms of mRNA concentrations $X_1$ and $X_2$ of two coupled cells. (c) Time waveforms of mRNA concentrations $X_1$ and $X_2$ of two coupled cells under external stimulation of light. The mRNA concentrations $X_1$ and $X_2$ are synchronized to the light. Comparing (a)-(c) with those in Fig. 2, it can be confirmed that the proposed model can reproduce generations of the typical circadian rhythms.

Future problems include: (a) analysis of large scale gene networks, (b) hardware implementation, and (c) comparison of hardware cost with previous models. This work was partially supported by JSPS KAKENHI Grant Number 15K00352.

References


Space-Time analogy in delay systems
for chimera states and Reservoir Computing

L. Larger †, B. Marquez †, B. Penkovsky †, M. Jacquot †, Y. K. Cheombo † and D. Brunner †

Abstract—Space-Time analogy for nonlinear delay dynamics was proposed in the early 1990s, as a tool to visualize and interpret complex motions in the time domain only. Beyond the visualization tool, this analogy was recently pushed further as a conceptual argument to support the emulation, through nonlinear delay dynamics, of traditional spatio-temporal dynamics such as the ones provided by neural networks or by paradigmatic models like network of coupled Kuramoto oscillators. The contribution will report on experimental and analytical results obtained for both Reservoir Computing and chimera states, and showing the powerful capability of delay system to emulate features usually attributed to network of oscillators.

1. Introduction

In the field of complex systems, network of oscillators are representing a widely used paradigm in order to numerically explore the dynamical features issued from high dimensional systems. From a more applied perspective, such studies are expected to provide insights for the understanding of real world complex systems, whether natural ones such as the brain, or technological ones such as power grid networks, among many other examples. For simplifications issues, partly motivated by more tractable problems addressed through numerical or theoretical approaches, perfectly identical oscillator networks have been also explored along the same research direction. Symmetry effects then are expected to come with more importance, however losing partly relevance with respect to realistic situation where small differences or variations are always experienced within a real world network of dynamical nodes.

We have recently contributed in two different research topics related to the oscillator network theory, through an unconventional virtual emulation of a network of oscillators by delay differential equations. Thanks to this virtual emulation, one has access to a network of rigorously identical oscillator, with strictly identical coupling. Moreover, due to the ease of physical implementation of delay systems, whether in electronic of photonics, such an approach could represent a very interesting new paradigm for the study oscillator networks. After a brief explanation of the theoretical concepts supporting the virtual emulation of oscillator networks through delay dynamics and thus the actual relevance of a space-time analogy for delay systems, we will report on the two topics through which we have tested this relevance : The study of chimera states occuring in identically coupled oscillators in a network, and the design and implementation of a novel neural network-based computing concept (Reservoir Computing).

2. Space-time analogy and impulse response modeling

We assume the system under study is belonging to the general class of Mackey-Glass or Ikeda delay dynamical systems [1, 2], where the dynamical variable is named \( x(t) \), and the involved delay is \( \tau_D \). In such systems, one can conceptually split the feedback systems into two sub-systems interacting circularly one with the other [3]: A linear dynamical part from which the dynamical variable \( x(t) = \mathcal{F}^{-1}[X(\omega)] \) is obtained (\( \mathcal{F}^{-1} \) stands for the inverse Fourier Transform), and which is modeled by a linear filtering \( H(\omega) \) in the Fourier domain; And an adiabatic (i.e. instantaneous, without any dynamical effect) nonlinear delayed part represented by the feedback signal \( z(t) = \mathcal{F}^{-1}[Z(\omega)] = f_{NL}[x(t - \tau_D)] \). With such an assumption, the modeling of the dynamics in the Fourier domain reduces to the very simple following equality:

\[
X(\omega) = H(\omega) \cdot Z(\omega)
\]  

From the previous equation, and invoking the conversion rules from the Fourier domain to the time domain (e.g. \( \mathcal{i} \omega X(\omega) \rightarrow \text{dx/dt} \)), one generally derives the delay differential equation ruling the dynamics of \( x(t) \) solely (without \( z(t) \) which can be now replaced by its definition depending on \( x(t - \tau_D) \) only) in the time domain. Such a derivation is however subjected to the knowledge of the exact filtering profile for \( H(\omega) \), typically in the form of a fraction of polynomials. The simplest form for such a filtering profile is the one of a low pass filter \( (H(\omega) = (1 + i \omega \tau)^{-1}) \), which results in the typical first order scalar delay differential equation as concerned with the Mackey-Glass or Ikeda models,

\[
\tau \frac{dx}{dt}(t) = -x(t) + f_{NL}[x(t - \tau_D)].
\]  

Such models, despite their large interest and the many publications reporting on their complex dynamical behaviors, however represent a very specific sub-class only of the class of problems modeled in the Fourier domain by Eq.(1).
Each kind of linear Fourier filter thus leads to a new sub-class of delay differential, as exemplified by the bandpass filtering case our group has studied since the early 2000 [4], and which has revealed many new dynamical phenomena (chaotic breathers, Neymarc-Sacker bifurcation, single period periodic motion, chimera states) compared to the most widely studied low-pass case. The simplest bandpass type filter is \( H(\omega) = \frac{i \omega \theta}{(1 + i \omega \theta)(1 + i \omega \tau)} \), where \( \tau \) and \( \theta \) are the characteristic time scales determining the high and low cut-off frequencies \( f_h = (2\pi \tau)^{-1} \) and \( f_l = (2\pi \theta)^{-1} \) respectively. Under this filtering model assumption, an integro-differential delay equation can be deduced:

\[
\frac{1}{\theta} \int_{h_0}^{h} x(\xi) \, d\xi + x(t) + \tau \frac{dx}{dt}(t) = z(t) = f_{NL}[x(t - \tau D)].
\]

(3)

Beyond these many particular cases derived for each new Fourier filtering profile, one can keep in the time domain the generality offered by the Fourier domain through Eq.(1), however losing the convenient and widely preferred differential equation description for the time. The direct conversion of Eq.(1) indeed results in a convolution product (thus a “global” integral representation of the dynamics, instead of the local one provided by a differential equation):

\[
x(t) = \int_{-\infty}^{\infty} h(t - \xi) \cdot f_{NL}[x(\xi - \tau D)] \, d\xi,
\]

(4)

where \( h(t) \) is the well known (causal) impulse response of the linear filter defined as the inverse Fourier transform of the Fourier filtering function \( H(\omega) \). Re-writing such an integral representation of the dynamics with some specific features known for delay equations, e.g. the actual infinite dimensionality of such dynamics because of the functional nature of its initial conditions (e.g. a function of time \( x_0(t) \) defined over the time interval \( t \in [-\tau_D, 0] \)), one can obtain the following expression [5]:

\[
x_n(\sigma) = x_{n-1}(\sigma) + \int_{\sigma-1}^{\sigma+\gamma} h(\sigma + \gamma - \xi) \cdot f[x_{n-1}(\xi)] \, d\xi,
\]

(5)

where the time is decomposed as follows, \( t = (n\eta + \sigma)\tau_D \), with \( n \in \mathbb{N} \) and with \( \eta \) being a constant close to unity, \( \eta = 1 + \gamma \) (with \( \gamma = o(\tau/\tau_D) \)), thus reflecting the time delay iteration process inherent to delay dynamics. From the previous equation which is rigorously derived analytically, one can clearly make a new physical interpretation in terms of network of coupled oscillators for the space-time analogy earlier proposed for delay equations [6]: The amplitude \( x_n(\sigma) \) of any oscillator corresponding to a virtual position \( \sigma \in [0, \eta] \), is dynamically ruled from the same amplitude \( x_{n-1}(\sigma) \) at one time delay earlier (iteration from \( (n - 1) \) to \( n \), resulting in a discrete time dynamics), with a modification ruled by the integral term. This integral term appears as a nonlinear coupling of the continuosly distributed neighboring oscillators at positions \( \xi \) around \( \sigma \), the impulse response \( h \) playing the role of a coupling coefficient.

In the next sections, we will illustrate for two particular situations, how such a space-time analogy was recently used on the one hand to discover the existence of chimera states in delay dynamics [5], and on the other hand to demonstrate the processing efficiency of delay systems when they are replacing the dynamics of a neural network to perform Reservoir Computing [7].

3. Chimera states in delay systems

Chimera states have been discovered numerically in 2002 by Kuramoto [8], while exploring the emergence of symmetry breaking solutions exhibited by network of coupled identical oscillators. In the case of long range (non local) coupling conditions, particular sustained solutions were observed, in which the whole network appears to be structured into sub-networks of congruent solutions within a sub-network, but incongruent between sub-networks. Chimera states have attracted lots of interest because of their non-intuitive features corresponding to symmetry breaking solutions within a network constructed with perfect symmetry. In the case of a network of phase oscillators, it can be found under appropriate coupling offset phase and coupling radius, that parts of the network exhibit fully synchronized oscillators whereas other parts show totally desynchronized ones. Both regions appear to coexist in a stable way within the whole network. After their first discovery, one had to wait 10 years until experimental observation of chimera states could be achieved in 2012. Two independent papers in two different fields, optics and chemistry, reported the experimental formation of such chimera patterns. One in a spatio-temporal dynamics of the intensity profile of a light beam, and another in the volume of a reactor where a Belousov-Zabotinsky chemical reaction was prepared.

In 2013 [9], based on the assumption that delay systems can mimick some features of spatio-temporal dynamics, we reported the first numerical and experimental observation of chimera state within the virtual space-time representation of this infinite dimensional dynamics. Such a representation precisely highlights the discrete time evolution along a vertical axis when \( n \) is incremented every time delay iteration, of the virtual spatial domain amplitude distribution within each time delay \( \{x_n(\sigma) \mid \sigma \in [0, \eta]\} \). Under appropriate parameter condition and delay dynamical model, the corresponding functional \( x(\sigma) \) evolving over time \( n \) was clearly exhibiting the emergence of a well structured virtual space along \( \sigma \), with an alternance of quiet plateaux and chaotic-like oscillations, sustained within the “length” \( \tau_D \) of the virtual space as the discrete time \( n \) is growing.

Figure 1 shows an example of an experimentally recorded scalar time trace \( x(t) \) from a bandpass delay dynamical system. The time series was then cut according to properly chosen “spatial” intervals such that each for \( t \in [n\eta\tau_D; (n + 1)\eta\tau_D] \), we stack vertically the color encoded amplitudes \( x(t) \). From this representation, a particu-
Figure 1: Space-Time plot of an experimental 3-headed chimera solution emerging (n growing from bottom to top) from a background noise.

ilar pattern can be clearly viewed and identified as a chimera state.

Thanks to the derivation of Eq.(5), interesting analogies and interpretation have been proposed in terms of coupling distance and its influence on the chimera existence and feature [5]. Since the coupling function between distant virtual oscillators turns out to be determined by the impulse response profile $h(\xi)$, one can carve such a function directly through the Fourier filter used in the delayed feedback loop. Work is in progress to demonstrate, via the convolution product description involving the impulse response, why chimera state can be found (or are not stable) in low pass delay dynamics, whereas they have been indeed observed for bandpass delay dynamics. One can notice for example, that a direct consequence of a bandpass filter compared to a low pass, is to extend the equivalent coupling range in terms of network of oscillators, through a broader impulse response.

Chimera states reveals deterministic organization of complexity within high dimensional dynamical systems. They correspond to the emergence of spontaneous complex dynamics in an autonomous way, in the sense there is no information provided by the external world of the dynamics, except for the initial noisy background from which chimera appears. In the next section, similar spatio-temporal features of high complexity delay dynamics will be reported, however in a strongly non autonomous way. Indeed, we will report on the processing capability of delay dynamics while they are subject to large amplitude external forcing coming from the information signal to be processed.

4. Reservoir Computing (RC) with delay systems

The concept of RC [10, 11] is derived from recurrent neural network (RNN) approaches, however simplifying extremely the learning phase of the computational steps. The latter indeed represents traditionally a very critical issue in standard RNN, because the optimal set of coupling parameters is very difficult to determine by a learning procedure, particularly when they concern many sets of such connectivity strength, the ones of the input and output layers, and the ones of the internal connectivity defining the network structure itself. RC considers that the output connectivity, also called the read-out or output layer only, needs to be learnt. The two other sets of connectivity coefficients do not need critical optimization, and they can be thus simply chosen at random for example. Such a simplification transforms the learning phase into a very simple, very efficient, very fast, and always converging solution. Beyond this surprising simplification, RC has moreover shown surprising computation accuracy, with comparable results, and sometimes even better ones, compared to traditional neural network computing.

More recently, RC has reached another important step forward through its successful hardware demonstration [12], moreover with an initially unexpected structural solution for the so-called Reservoir: the usual network of interconnected nodes was physically realized through the internal complexity of a delay dynamical system. As illustrated in the theoretical arguments of Section 2, delay dynamics can provide qualitatively similar complexity features compared to spatio-temporal dynamics such as a network of neurons. In the present section, we will again take the opportunity of the unusual modeling of delay dynamics through Eqs.(4) and (5), in order to analytically derive a rigorous correspondence between a delay dynamics seeded by a time division multiplexed input information, and a network of interconnected nodes excited by an input information through the usual input layer.

A key concept in the use of a delay system to emulate a neural network, is to consider the dynamical nodes of the network as being temporal positions within the time interval corresponding to the delay. One needs then to re-define the time variable $t$, so that it can reflect the emulation of a virtual spatial position $\sigma \in [0, \eta]$, which is updated in time each round trip of the signal in the delayed feedback loop, i.e. each time delay $\tau_D$. Such an approach indeed reveals the intrinsic multiple time scale feature of a delay dynamics, the fast time scale $\tau$ related to the high cut-off frequency $f_h$, and the slow time related to the delay $\tau_D$: $t = (\sigma + n) \cdot \eta \tau_D$.

If one then assumes that the virtual nodes correspond to sampled positions $\sigma_k = k \delta \tau / \tau_D$, the number of virtual nodes in the delay dynamics amounts to $K = \tau_D / \delta \tau$. Addressing each of these nodes with an input vector $\mathbf{u}(n) \in \mathbb{R}^D$ is achieved, as already stated, through a standard time division multiplexing technique. Distributing “randomly” each vector component of $\mathbf{u}(n)$ onto each of the $K$ virtual nodes of the delay dynamics, is an operation typically performed according to a so-called input connectivity matrix. From Eq.(4), one can arrange the integration interval for the convolution product so that the node amplitude $x_k \equiv x_{\sigma_k}$ at time $n$ can be expressed as an update of the amplitude of the same node, but at time $(n - 1)$, i.e. $x_k(n - 1)$. Taking
also into account that the dynamics is seeded by the input information to be processed, one obtains:

\[ x_k(n) = x_k(n-1) + \int_{\sigma_k-\tau_0}^{\sigma_k} h(\sigma - \sigma_k) \times \]

\[ f_{\text{nl}}[x_k(n-1) + \rho \cdot u'(n-1)] \, d\sigma. \quad (6) \]

The latter expression reveals in a rigorous way the analogy of delay-based RC with the original Echo State Network approach as proposed in [10].

The output layer consists also in a matrix multiplication, corresponding physically to a circular convolution operated on the response signal \( x_k(n) \) and involving the Read-Out matrix \( W^R = \{w^R_{mk}\} \in \mathbb{R}^M \times \mathbb{R}^K \). The computed output is a vector \( y(n) \in \mathbb{R}^M \), which is the expected calculation result obtained from the input information \( u(n) \):

\[ y_m(n) = \sum_{k=1}^{K} w^R_{mk} x_k(n). \quad (7) \]

The Read-Out matrix \( W^R \) is practically the solution of a ridge regression problem minimizing the error of the output vector considering a set of known pairs of answer / response \( (u(n); y(n)) \). This ridge regression step precisely corresponds to the learning phase of such a delay-based RC.

This delay-based RC concept was practically implemented recently by different authors, with very successful computational performances. A classical speech recognition problem was for example performed experimentally [7, 12], with record word error rate (WER) down to 0% for a clean spoken digit database, thus achieving state of the art performances.

5. Conclusion

We have reported an original writing of a delay dynamics through a signal processing approach, involving a convolution product description instead of the usual delay differential equation. The temporal impulse response attached to the linear Fourier filter involved in the delayed feedback oscillator loop, was revealed as a key physical ingredient for such a convolution product description. This representation was particularly useful to identify the space-time analogy involved in two currently investigated research topic for delay dynamics: Chimera states, and delay-based Reservoir computing. Our analytical derivation has shown its relevance in the interpretation of these two recent successful achievements. Work is in progress to further develop this theory, in order to better understand chimera states, as well as to optimize the computational capabilities of delay-based RC.

Acknowledgments

This work was supported by the ANR project BiPhoProc (contract ANR-14-OHRI-0002-02), and the Labex ACTION program (contract ANR-11-LABX-01-01). BM and BP acknowledge the support of the Region FrancheComté. 

References


Fast information processing by using fast transient response in a semiconductor laser with strong optical injection

Kazutaka Kanno† and Masatoshi Bunsen†

†Department of Electronics Engineering and Computer Science, Fukuoka University
8–19–1 Nanakuma, Johnan-ku, Fukuoka, Japan
Email: kkanno@fukuoka-u.ac.jp, bunsen@fukuoka-u.ac.jp

Abstract—Reservoir computing (RC) is a machine learning paradigm based on information processing in the human brain. Recently, RC based on a semiconductor laser with time-delayed optical feedback has been proposed. In this scheme, fast transient response (a fast relaxation time) of the laser is necessary for fast information processing. By using numerical and linear stability analysis, in this study, we show that a semiconductor laser with strong optical injection can produce fast transient response. We also numerically demonstrate that RC based on a semiconductor laser with optical feedback and strong optical injection enables fast information processing.

1. Introduction

Delay-based Reservoir Computing (RC) has been proposed as an information processing method using time-delayed dynamical systems [1–4]. RC is a machine-learning paradigm that can process empirical data and is inspired by the way that the brain processes information [5]. A time-delayed dynamical system is treated as a virtual network, where nodes are considered by temporally dividing feedback delay with a small interval \( \theta \), which is called a node interval. When an input signal is injected into the time-delayed dynamical system, the system produces a transient response and virtual node states can be obtained from the response. The output of RC is given by a weighted linear combination of the virtual node states, where the weights are decided in a training procedure.

The node interval \( \theta = 0.2 \cdot T_{ro} \) has been selected in some literatures [1,3], where \( T_{ro} \) is the characteristic time scale of the system’s relaxation oscillations (ROs). For smaller \( \theta \), dynamical systems cannot respond to an input signal. For larger \( \theta \), the connectivity among the virtual nodes is lost. We can increase the information processing speed in RC using a small \( \theta \) because a time for processing an input data point corresponds to \( N \cdot \theta \), where \( N \) is the number of nodes. Therefore, time-delayed dynamical systems with fast ROs are required to enable a smaller \( \theta \). In RC using a semiconductor laser with time-delayed optical feedback, \( \theta = 0.2 \) ns has been achieved [3]. The laser’s RO frequency corresponding to a few GHz enables the small \( \theta \).

It has been studied that strong optical injection can enhance the RO frequency of an optically injected semiconductor laser [6]. In this study, we show that the node interval \( \theta \) can be reduced in RC based on a semiconductor laser with time-delayed optical feedback and optical injection when the optical injection strength is strong.

In addition, we investigate the dependence of the performance of our RC on the injection strength and the initial optical frequency detuning between the laser with optical injection and the injected light. It has been known that the initial detuning affects synchronization properties and consistency in coupled semiconductor lasers [7,8]. It is expected that the initial detuning also affects transient dynamics in the laser and the information processing performance of RC.

2. Numerical simulation method for RC with semiconductor lasers

2.1. Delay-based RC method

Figure 1 shows the schematic diagram of our RC system. The system is composed of an input layer, a reservoir, and an output layer. The reservoir is a semiconductor laser (called the response) with time-delayed optical feedback and optical injection from another laser (called the drive). \( N \) virtual nodes are considered by temporally dividing the feedback delay time \( \tau \). A time interval for the division is \( \theta \) and is called the node interval.

In the input layer, time-discrete input data \( s_n \) (\( n \) is the discrete time) are multiplexed by a temporal mask signal. The mask signal is a step waveform which has a period \( \tau \). The step interval of the mask signal is equal to the node interval \( \theta \). The mask values are randomly selected from the
values $[-1, -0.6, 0.6, 1.0]$.

A weighed linear combination of virtual node states is calculated in the output layer and is the output of the reservoir. The output $y(n)$ for the $n$-th input data is given by the following equation,

$$y(n) = \sum_{j=1}^{N} w_{j} x_{j}(n),$$  \hspace{1cm} \text{(1)}

where $x_{j}$ are the node states and $w_{j}$ are the weights for $j$-th node. The node states $x_{j}$ are sampled at the center of the node interval $\theta$ in the temporal output of the response laser. The weights $w_{j}$ are trained by minimizing the mean-square error between the target function $\hat{y}(n)$ and the reservoir output $y(n)$ as follows,

$$\frac{1}{N_{e}} \sum_{j=1}^{N_{e}} (y(n) - \hat{y}(n))^{2} \rightarrow \text{min},$$  \hspace{1cm} \text{(2)}

where $N_{e}$ is the number of input data for training.

To evaluate the performance of our RC scheme, we use the Santa-Fe time-series prediction task [1–4]. The aim of the task is to perform single-point-prediction of chaotic time-series. The time-series is generated from a far-infrared laser. We use 3,000 points for training and 1,000 points for testing.

The performance of the prediction task is quantitatively evaluated by using the normalized mean-square error (NMSE) as follows,

$$\text{NMSE} \equiv \frac{1}{N_{e}} \frac{\sum_{j=1}^{N_{e}} (y(n) - \hat{y}(n))^{2}}{\sigma^{2}},$$  \hspace{1cm} \text{(3)}

where $N_{e}$ is the number of input data in the test procedure. $\sigma$ is the standard deviation of $\hat{y}(n)$. The NMSE represents the difference between the target $\hat{y}(n)$ and the output $y(n)$ of RC, and a NMSE close to zero indicates a low prediction error.

2.2. Numerical model of the response laser

We consider two unidirectionally coupled semiconductor lasers, which are called drive and response. The drive laser solitary operates so that the laser shows temporally constant intensity. The constant output of the drive laser is unidirectionally injected into the response laser, which also have time-delayed optical feedback. We consider the dynamics of the response laser since the output of the drive laser is constant. The rate equations for the response laser are written as follows [9]:

$$\frac{dE_{r}(t)}{dt} = \frac{1 + i\omega}{2} \left[ \frac{G_{N}(N_{r}(t) - N_{0})}{1 + |E_{r}(t)|^{2}} - \frac{1}{\tau_{p}} \right] E_{r}(t)$$

$$+ \kappa E_{r}(t - \tau) \exp[-i\varphi]$$

$$+ \sigma(1 + \Delta \omega(t)) \lambda_{d} \exp[i\Delta \omega t] + \xi(t),$$  \hspace{1cm} \text{(4)}

$$\frac{dN_{r}(t)}{dt} = J_{r} - \frac{N_{r}(t)}{\tau_{s}} - G_{N}(N_{r}(t) - N_{0})|E_{r}(t)|^{2},$$  \hspace{1cm} \text{(5)}

where $E$ is the slowly varying complex electric field amplitude and $N$ is the carrier density. The subscripts $d$ and $r$ represent the drive and response lasers, respectively. $G_{N}$ is the gain coefficient, $N_{0}$ is the carrier density at transparency, $\alpha$ is the linewidth enhancement factor, $\tau_{p}$ is the photon lifetime, $\tau_{s}$ is the carrier lifetime, and $J_{r}$ is the injection current of the response laser. The injection current is given by $J_{r} = 1.05J_{th}$, where $J_{th}$ is the injection current at the lasing threshold. $\omega_{0}$ is the angular optical frequency of the response laser. These parameter values are set to the same as in [8].

Optical feedback is related to the second term in the right hand side of Eq. (4). $\kappa$ and $\tau$ in the term are the feedback strength and the feedback delay time, respectively. The delay time is given by the product of the number of nodes $N$ and the node interval $\theta$ in RC. In this study, the number of nodes is two hundreds and the node interval is varied, which results in $\tau = 200 \cdot \theta$. $\varphi$ is the feedback phase and fixed at zero for simplicity.

The third term in the right hand side of Eq. (4) represents optical injection, through which an input signal is injected. $\sigma$ is the optical injection strength. $\Delta \omega(t)$ is the angular optical frequency detuning between the drive and response lasers and given by $\Delta \omega(t) = 2\pi \Delta f_{inj}$, where $\Delta f_{inj}$ is the initial optical frequency detuning and is changed in our study. $A_{d}$ is the constant electric field amplitude of the drive laser and is calculated from steady state solutions of a solitary laser. In the calculation of the solutions, the injection current $J_{d}$ of the drive laser is $1.30J_{th}$ and other parameter values are the same as the response laser.

3. Numerical results on RC for chaotic time-series prediction task

3.1. Dependence of RC performance on node interval

We numerically show that strong optical injection induces fast oscillations in the transient response of the laser, which results in a broad probability distribution in node states. We also investigate the dependence of the NMSE on the node interval $\theta$ for weak and strong injection strengths. It is shown that a minimum NMSE is obtained at a smaller $\theta$ for strong optical injection comparing with weak one.

We firstly show the temporal waveforms of the intensity $I(t) = |E_{r}(t)|^{2}$ in the response laser when the input signal shown in Fig. 2(a) is injected. Figures 2(b) and 2(c) show the temporal waveforms when the node interval $\theta$ is 0.03 ns. The optical injection strengths are $5 \text{ ns}^{-1}$ and $40 \text{ ns}^{-1}$ for Figs. 2(b) and 2(c), respectively. The red circles represent the node states. Since the number of nodes is two hundreds, the feedback delay time is $\tau = 200 \theta = 6 \text{ ns}$. The initial optical frequency detuning $\Delta f_{inj}$ is fixed at 0 GHz. The feedback strengths $\kappa$ are $1 \text{ ns}^{-1}$ and $22 \text{ ns}^{-1}$ for (b) and (c), respectively.

For the weak injection strength in Fig. 2(b), oscillations in the waveform are slow in comparison with the input sig-
Figure 2: (a) A temporal waveform of a masked input signal. (b), (c) The temporal waveforms of the response intensity. The injection strengths $\sigma$ are 5 ns$^{-1}$ and 40 ns$^{-1}$ for (b) and (c), respectively. The red circles represent node states. The node interval $\theta$ is 0.03 ns. (d), (e) Probability distributions of the node states. The distributions (d) and (c) correspond to (b) and (c), respectively.

On the other hand, fast oscillations corresponding to the input signal are observed in the waveform shown in Fig. 2(c) for the strong optical injection. The fast oscillations result from fast ROs due to the strong optical injection [6]. When the injection strength is weak, the laser cannot respond to the input signal since the RO frequency of the laser is slower than the modulation frequency of the input signal.

When oscillations in the transient response is slow, as shown in Fig. 2(b), the width of the probability distribution of node states becomes narrow. Figures 2(d) and 2(e) show the probability distributions of node states corresponding to Figs. 2(b) and 2(d), respectively. It is found that the narrow distribution is obtained for the weak injection strength in comparison with the strong one. The narrow distribution indicates that the variety of node states is not rich, which causes reduction of the RC performance. The NMSEs for the weak and strong injection strength are 0.148 and 0.038, respectively. Thus, the better performance is obtained when node states have a broad distribution.

We show the dependences of the NMSE on the node interval $\theta$ for strong and weak optical injection and investi-gate the interval at which the minimum value of the NMSE is obtained. Figures 3(a) and 3(b) show the dependences when the optical injection strengths are weak ($\sigma = 5$ ns$^{-1}$) and strong ($\sigma = 40$ ns$^{-1}$), respectively. The minimum NMSEs are obtained at $\theta = 0.18$ ns and 0.03 ns for the weak and strong injection strength, respectively. This result indicates that strong optical injection enables us to use a small $\theta$.

3.2. Dependence of RC performance on coupling parameters

Strong optical injection can induce injection locking in coupled semiconductor lasers. In RC with the laser, injection locking is necessary for consistency and keeping the laser stable when the injected light into the laser is not modulated. The injection strength $\sigma$ and the initial optical frequency detuning $\Delta f_{ini}$ are important parameters for injection locking. In this section, we show injection locking region in the two-dimensional parameter space and investigate the dependence of the NMSE on the two parameters.

Injection locking can be identified by the optical frequency detuning between the drive and response lasers under coupling. The detuning $\Delta f_c$ is given by the following equation,

$$\Delta f_c = \Delta f_{ini} + \frac{\Delta \phi(t) - \Delta \phi(t - T_d)}{2\pi T_d},$$

(6)

where $\Delta \phi(t)$ is the phase difference between the drive and response lasers and is given by $\phi_d(t) - \phi_r(t)$. $\phi(t)$ is the phase of the complex electric field amplitude $E(t)$ and the subscripts $d$ and $r$ represent the drive and response lasers, respectively. Since the optical output of the drive laser is temporally constant, $\phi(t)$ is also temporally constant, which results in $\Delta \phi(t) = -\phi(t)$. $T_d$ is a time for the convergence of $\Delta f_c$ and $T_d = 5000$ ns is used in our numerical simulation. $\Delta f_c = 0$ indicates the match of the optical frequencies of the drive and response lasers, that is, injection locking.

Figure 4(a) shows injection locking region on the two-dimensional space of the injection strength $\sigma$ and the ini-
Figure 4: (a) Two-dimensional map of $|\Delta f_c| \leq 0.1$ GHz on the parameter space of the injection strength $\sigma$ and the initial optical frequency detuning $\Delta f_{ini}$. A region with $|\Delta f_c| \leq 0.1$ GHz is shown by the gray color. (b) NMSE corresponding to (a).

Figure 5: Two-dimensional map of the conditional Lyapunov exponent $\lambda_c$ corresponding to Fig. 4.

The dependence of the NMSE on $\Delta f_{ini}$ when the injected light into the response laser is not modulated. In the gray region, $|\Delta f_c| \leq 0.1$ GHz is obtained and $\Delta f_c$ close to zero results from injection locking. An asymmetric property for $\Delta f_{ini}$ is observed and negative detuning makes it easy to achieve injection locking. This property results from a large $\sigma$ of the laser.

The two-dimensional map of the NMSE corresponding to Fig. 4(a) is shown in Fig. 4(b). White and black colors in the map represent small and large NMSEs, respectively. A large NMSE represented by the black color is obtained outside the gray region shown in Fig. 4(a). It indicates that the performance of RC deteriorates when injection locking does not occur. A small NMSE represented by the white color is obtained near the boundary between the gray and white regions shown in Fig. 4(a) inside the gray region. It’s worth noting that a small NMSE is not obtained at the side of negative detunings but is obtained at the side of positive ones.

The dependence of the NMSE on $\Delta f_{ini}$ can be explained using the conditional Lyapunov exponent [8] for a synchronized solution between the response laser and its auxiliary system [10]. The conditional Lyapunov exponent is an exponential convergence (growth) rate of perturbations to a synchronized solution and a negative exponent indicates that the synchronized solution is stable. A negative value of the exponent is necessary for consistency. The exponent close to zero is also required.

Figure 5 shows the conditional Lyapunov exponent $\lambda_c$ corresponding to the two-dimensional map of $\sigma$ and $\Delta f_{ini}$ when the injected light into the response laser is not modulated. From comparing to Fig. 4(a), the region with $\lambda_c \leq 0.0$ corresponds to the injection locking region shown by the gray region. $\lambda_c$ around the boundary with $\lambda_c = 0.0$ shows gradual changes at the side of positive detunings and sudden changes at the side of negative ones. Thus, a negative $\lambda_c$ close to zero is obtained at the side of positive detunings, which results in a small NMSE shown in Fig. 4(b).

4. Conclusion

In this study, we numerically demonstrated RC with a semiconductor laser with time-delayed optical feedback and strong optical injection. The performance of RC was quantitatively evaluated using the chaotic time-series prediction task. It was shown that strong optical injection induces fast oscillations in the transient response of the laser, which enables a small node interval in RC. We also investigated the dependence of the RC performance on the injection strength and the initial optical frequency detuning. Strong optical injection induces injection locking in the laser and high performance of RC can be obtained in an injection locking region on the two-dimensional parameter space. A positive detuning near the boundary of injection locking is also required for high performance of RC.

References

Experiment on reservoir computing using consistency of a semiconductor laser

Yoma Kuriki, Joma Nakayama, Kosuke Takano, and Atsushi Uchida

Department of Information and Computer Sciences, Saitama University, 255 Shimo-Okubo, Sakura-ku, Saitama City, Saitama, 338-8570, Japan
Emails: s15mm310@mail.saitama-u.ac.jp, auchida@mail.saitama-u.ac.jp

Abstract—Reservoir computing is an efficient approach for processing time dependent information, such as time series prediction and speech recognition. We experimentally investigate reservoir computing based on the consistency of a semiconductor laser subjected to optical feedback and injection. We investigate the performance of a time-series prediction task by changing an input temporal mask. We evaluate the influence of noise when the temporal waveforms are averaged.

1. Introduction

Reservoir computing (RC) has been intensively investigated as information processing inspired by the brain [1-6]. RC is one type of recurrent neural networks, and the weight of input and the connections in a network (called reservoir) are fixed and only the output weights are trained by machine learning. Computational complexity for learning can be reduced in RC. In 2011, an implementation of RC using one nonlinear element with time-delayed feedback has been proposed [2]. The network is constructed from many virtual nodes of the output of the nonlinear element within the time-delayed feedback loop (see Fig.1).

![Figure 1: Schematics of RC using a nonlinear element with time-delayed feedback.](image1)

For delay-based RCs, an input signal is multiplied with a mask signal as preprocessing to increase the complexity of the reservoir outputs. The mask signal is used to map an input signal into a high dimensional space. A binary random signal has been used as the mask signal in many implementation of RC [2, 5, 7]. In addition, a six-level random mask [8] and chaos mask signals [9] are used to improve the performance of RC.

In this study, we experimentally implement RC using a semiconductor laser with optical feedback and injection. We evaluate the performance of RC by using a time-series prediction task. We investigate the performance of RC when the mask signal is varied. We also evaluate the performance of RC when the output waveforms of the response laser are averaged to reduce the effect of noise.

2. The concept of RC and experimental setup

![Figure 2: Schematics of RC using two semiconductor lasers.](image2)

The use of an optical device for the reservoir results in high-speed implementation of RC at a rate of several GHz. Furthermore, semiconductor lasers with optical injection have a property called consistency [3]. Consistency is defined as reproducibility of the outputs in the response system driven repeatedly by an identical drive signal. Consistency property is required for the implementation of RC.
The reservoir is constructed using two semiconductor lasers (called drive and response lasers, see also Fig.3). The modulation signal is generated from an arbitrary waveform generator and sent to a phase modulator. The output of the drive laser is modulated by the phase modulator. The modulation signal consisting of the input and the mask signals is converted into an optical signal, and injected into the response laser. The output of the response laser is observed by using a digital oscilloscope. The output of the response laser is determined by the injection signal and the optical feedback signal with time delay.

The outputs within the feedback loop are used as virtual node states of the reservoir. The virtual nodes $x_i(i = 1, 2, ..., N)$ are distributed within the output of the feedback loop with the interval of $\theta$. In the output layer, the output signal $y(n)$ is calculated from the weighted sum of the virtual node states in the reservoir as follows:

$$y(n) = \sum_i w_i x_i(n)$$

(1)

The weights $w_i$ are optimized by the linear least-squares method with the training data. We evaluate the performance of RC by comparing the output signal $y(n)$ with the input data. We set the parameter values in the experiment as follows: $\tau = 35.4$ ns, $\theta = 0.2$ ns, and $N = 177$.

$$NMS E = \frac{1}{L} \sum_{n=1}^{L} (y(n) - Y(n))^2 / \text{var}(y)$$

(2)

Where $L$ is the index of the input signal, $y(n)$ is the predicted signal by using RC and $Y(n)$ is the actual input signal at the next time step.

Figure 4(a) shows the result of the time-series prediction using a binary random mask signal, which is shown in Fig. 4(b). The upper waveform of Fig. 4(a) is the input data, and the lower waveform is the predicted signal by using the RC. The error of $NMS E = 0.216$ is obtained for the binary random mask signal.

3. Investigation of RC performance

3.1. Use of binary random mask signal

We investigate the performance of the RC using time-series prediction task. In this task, RC predicts the value of an input signal at the next time step. We use a chaotic time series of the Santa Fe time-series data set as an input signal [10]. The data consists of 4,000 points, and we use 3,000 points for training and remaining 1,000 points for testing. The normalized mean square error ($NMS E$) is used to evaluate the performance of the time-series prediction task as follows.

We investigate the performance of RC using a different mask signal in the input layer. We use six-level mask [1, 0.84, 0.3, -0.3, -0.75, -1] and the values between [1, 0.84] are randomly fluctuated, as shown in Fig. 5(b). Figure 5(a) shows the result of the time-series prediction using the improved six-level mask. The error of $NMS E = 0.139$ is obtained in Fig. 5(a). The prediction error using this improved six-level mask is smaller than that using the binary mask signal. We speculate that the random fluctuation of the mask signal with large amplitudes results in complex response of the reservoir, and this effect could improve the performance of the prediction task.

3.3. Performance of RC by averaging temporal waveforms

We average temporal waveforms of the response laser output used for the calculation of virtual node states in order to reduce the effect of noise. The performance of the time-series prediction task is degraded by the influence of the noise in the semiconductor laser and the detection noise. We obtain the temporal waveforms of the response laser output repeatedly using the same modulation signal.

Figure 3: Experimental setup of the reservoir with two semiconductor lasers.

Figure 4: Temporal waveforms of (a) the original input signal and the predicted signal, and (b) the binary random mask signal.

Figure 5: The output signal of the response laser is observed by using a digital oscilloscope. The output of the response laser is determined by the injection signal and the optical feedback signal with time delay.

The reservoir is constructed using two semiconductor lasers.
The noise is reduced by averaging the obtained temporal waveforms.

Figure 6 shows the result of the time-series prediction task by averaging 25 temporal waveforms of the response laser output in the case of the binary mask signal. The prediction error of $NMS E = 0.0279$ is obtained and significant improvement of $NMS E$ is achieved.

We calculate the signal-to-noise ratio ($SNR$) as follows.

$$SNR = 10 \log_{10} \frac{var(S)}{var(N)}$$

(3)

Where $var(N)$ is the variance of the noise signal and $var(S)$ is the variance of the temporal waveforms of the response laser output with the input signal.

Figure 7 shows $NMS E$ and $SNR$ when the number of the waveforms used for averaging is changed. $NMS E$ decreases and $SNR$ increases as the number of the waveforms for averaging is increased. The values of $NMS E = 0.0279$ and $SNR = 21.1$ dB are obtained for 25 averaged temporal waveforms, compared with the values of $NMS E = 0.216$ and $SNR = 8.14$ dB for a single temporal waveform. This result indicates that the reduction of the noise signal by averaging can improve the performance of the prediction task (small $NMS E$).

4. Conclusions

We experimentally investigated reservoir computing based on the consistency of a semiconductor laser subjected to optical feedback and injection. We introduced a six-level mask signal with random fluctuations for large amplitudes, and we succeeded in improving the performance of the time-series prediction task. We reduced the effect of noise by averaging the temporal waveforms of the response laser output. The performance of the time-series prediction can be improved by reducing the amount of the noise.

Acknowledgments

We acknowledge support from Grants-in-Aid for Scientific Research from Japan Society for the Promotion of Science (JSPS KAKENHI Great Number JP24686010), and Management Expenses Grants from the Ministry of Education, Culture, Sports, Science and Technology in Japan.

References


[10] Santa Fe time-series prediction,
URL: http://wwwpsych.stanford.edu/ andreas/Time-Series/SantaFe.html
Period-One Nonlinear Dynamics of Semiconductor Lasers for Photonic Microwave Mixing

Yu-Han Hung1 and Sheng-Kwang Hwang1,2,*

1Department of Photonics, National Cheng Kung University, Tainan, Taiwan
2Advanced Optoelectronic Technology Center, National Cheng Kung University, Tainan, Taiwan
*Email: skhwang@mail.ncku.edu.tw

Abstract—Photonic microwave mixing has been considered a key functionality in radio-over-fiber systems adopting high microwave subcarrier frequencies for antenna remoting applications. Such a functionality enables microwave subcarrier frequency upconversion for wireless transmission in downlinks or downconversion for photodetection in uplinks through photonic approaches. In this study, an approach is proposed by taking advantage of the nonlinear wave mixing inside a semiconductor laser between an input optical signal carrying a microwave subcarrier and the period-one nonlinear dynamics invoked by the input optical signal.

1. Introduction

Period-one (P1) nonlinear dynamics in a semiconductor laser subject to continuous-wave (CW) optical injection have attracted much research interest not only for fundamental understandings of nonlinear dynamics and laser physics [1-7] but also for various technological applications in photonics and microwaves [8-24]. For example, by taking advantage of the self-sustained microwave oscillation of the laser intensity, the P1 dynamics have been proposed for photonic microwave generation [8-10, 12, 15, 16, 19-21, 24, 25]. While broadband frequency tunability can be achieved by simply adjusting the power and frequency of the optical injection, optical single-sideband modulation is so feasible as to mitigate microwave power fading over fiber distribution. In addition, by taking advantage of the multiple spectral components induced by the optical injection, the P1 dynamics have been proposed for optical frequency conversion [13]. While optical frequency down-, no-, and up-conversion can be simultaneously or individually achieved, the data modulation format, either amplitude modulation (AM), frequency modulation (FM), or phase modulation, can be maintained after conversion. Moreover, since the intensity and frequency of each spectral component depend on the injection level and frequency, the P1 dynamics have also been demonstrated for modulation format conversion between optical AM and optical FM [18] and from optical AM to microwave FM [11]. While different output modulation indices can be achieved by using different spectral components or different injection conditions, simultaneous optical frequency conversion is also possible. In addition, by adopting the intensity asymmetry between the oscillation sidebands, the P1 dynamics have been investigated for conversion from optical double-sideband modulation to optical single-sideband modulation [22]. Self-adaptation to changes in the operating microwave frequency is feasible, and stable operation under fluctuations of the injection level and frequency is achievable. Furthermore, the P1 dynamics have also been studied for photonic microwave amplification [23] by applying the red-shifted cavity resonance enhancement. The amplification can be achieved for a broad microwave range, up to at least 60 GHz, and for a wide gain range, up to at least 30 dB.

In this study, an approach is proposed for highly efficient photonic microwave mixing by taking advantage of the nonlinear wave mixing occurred inside a semiconductor laser between a microwave-modulated (MM) optical input and the P1 dynamics of the semiconductor laser invoked by the MM optical input. Photonic microwave mixing has been considered a key functionality in radio-over-fiber systems adopting high microwave subcarrier frequencies for antenna remoting applications, such as broadband wireless networks and electronic warfare systems. Such a functionality enables microwave subcarrier frequency upconversion or downconversion. Photonic approaches based on, for example, optical intensity modulators, optical phase modulators, and semiconductor optical amplifiers [26-32] provide various promising advantages, including broadband frequency tunability for either upconversion or downconversion, infinite isolation between microwave subcarriers and microwave local oscillators, and immunity to electromagnetic interference, which are difficult to achieve using electronic approaches. However, these photonic approaches typically suffer from significant power conversion loss of microwave subcarriers, require substantial power of electronic microwave local oscillators, experience considerable power loss of optical inputs, and need high-speed capability of photonic and/or electronic devices. The approach proposed in this study highly improves these performance characteristics and operating requirements.

2. Experimental Setup

Figure 1 presents a schematic of the experimental setup.
The proposed photonic microwave mixing system consists of a single-mode distributed-feedback semiconductor laser (Gooch & Housego AA0702), LD2. Under a bias current of 33 mA and a stabilized temperature of 25.6°C, the free-running LD2 oscillates at 193.345 THz with an optical power of 4.9 mW and a relaxation resonance frequency of about 10 GHz. An input optical carrier is generated by another single-mode distributed feedback semiconductor laser of a similar type, LD1, and is directed toward LD2 through a circulator. To excite the P1 dynamics, the frequency of the input optical carrier is detuned by \( f_i \) from the free-running frequency of LD2 through adjusting the temperature or bias current of LD1. In addition, the power of the input optical carrier is varied using a power adjuster consisting of an attenuator and/or an amplifier, and is measured at the output port of the circulator connected to the slave laser. To indicate the injection strength received by LD2, an injection ratio \( \xi_i \), defined as the square root of the power ratio between the input optical carrier and the free-running LD2, is used. A polarization controller aligns the polarization of the input optical carrier with that of LD2 to maximize the injection efficiency. An external modulator (EOspace AX-AV5-40) superimposes a microwave subcarrier at a frequency \( f_m \) from a microwave source (Agilent E8257D) on the input optical carrier. Data from a pattern generator (Anritsu MP2101A) are added onto the microwave subcarrier through an electronic microwave mixer. The output of LD2 is sent through a tunable optical bandpass filter (Alnair Labs BVF-200CL) to select the spectral components of interest before entering an optical spectrum analyzer (Advantest Q8384) and a microwave spectrum analyzer (Agilent N9030A PXA) following a 50-GHz photodiode (u2t Photonics XPDV2120R). For the bit-error ratio (BER) analysis, the photodetected signal is first downconverted to the baseband and next sent through an electrical low-pass filter before entering an error tester (Anritsu MP2101A).

3. Results and Analyses

To gain understanding of why the proposed mixing approach can perform as indicated above, let us first study the spectral features of a P1 dynamical state when LD2 is subject to a CW optical input at \((\xi_i, f_i) = (1.01, 20 \text{ GHz})\), shown as the black curve in Fig. 2. Not only the CW optical input regenerates at the offset frequency of 20 GHz, but also oscillation sidebands equally separated from the regeneration by \( f_0 = 30 \text{ GHz} \) sharply emerge. This characteristic of self-sustained microwave oscillation suggests that an optically injected laser at the P1 dynamics can work by itself as a microwave local oscillator, a photonic yet all-optical one, for photonic microwave mixing. In addition, \( f_0 \) is not limited by the laser intrinsic response and, in fact, can be broadly tuned from a few gigahertz to tens or even hundreds of gigahertz by simply adjusting \((\xi_i, f_i)\) [20, 22, 23]. Hence, as opposed to most other mixing approaches, no electronic microwave local oscillator is required, thus reducing system power consumption, and no high-speed photonic device, the laser here, is needed, thus relaxing high-frequency capability requirements. Note that, owing to the laser intrinsic noise, the microwave stability of the P1 dynamics is typically poor, on the order of 1 to 10 MHz [24, 25]. To demonstrate the best possible performance characteristics of the proposed mixing approach, a microwave stabilization scheme based on double locking [8] is adopted in this study, not shown in Fig. 1.

To demonstrate microwave mixing, an MM optical input at \( f_m = 35 \text{ GHz} \), as the red curve in Fig. 2 shows, with the optical carrier 30-dB stronger than both modulation sidebands, corresponding to an optical modulation depth of about 6%, is injected into LD2 at the same \((\xi_i, f_i) = (1.01, 20 \text{ GHz})\). As the blue curve in Fig. 2 presents, not only the MM optical input regenerates itself, including the optical carrier and both modulation sidebands, but also the optical carrier excites a P1 dynamical state at \( f_0 = 30 \text{ GHz} \) with key features closely similar to the one as the black curve shows. Since both modulation sidebands are too weak to frequency-lock the oscillation sidebands, nonlinear wave mixing between the excited P1 dynamical state and the regenerated MM optical input happens inside LD2 [22].
excited lower oscillation sideband and the regenerated lower modulation sideband using the tunable optical bandpass filter, as shown in Fig. 3, an optical output with two dominant tones is obtained, which are separated by \( f_m - f_0 = 5 \) GHz and which are 6-dB different in power, suggesting an improvement of the optical modulation depth as compared with that of the MM optical input. After photodetection, as Fig. 4 presents, this optical output results in a downconverted microwave subcarrier at 5 GHz with power amplification of 18 dB, suggesting a conversion gain, and with a 3-dB microwave linewidth of less than 1 Hz, the same as the input microwave subcarrier.

Figure 5 presents the BER analysis at a data rate of 1.25 Gb/s. The BER behavior of the downconverted microwave subcarrier is similar to that of its corresponding input microwave subcarrier, where a BER down to \( 10^{-9} \) is achieved. Owing to the conversion gain, a sensitivity improvement of about 9 dB is achieved, which is about one half of the conversion gain shown in Fig. 4. These results suggest that the quality of the data is mostly preserved after conversion.

4. Conclusion

This study investigates photonic microwave mixing by taking advantage of the nonlinear wave mixing occurred inside a semiconductor laser between a microwave modulated optical input and the P1 dynamics of the semiconductor laser invoked by the optical input. The laser works not only as a photonic microwave mixer but also as a photonic microwave local oscillator. Hence, no electronic microwave local oscillator is required and no high-speed semiconductor laser is needed. Owing to the improvement of the optical modulation depth after conversion, a conversion gain of 18 dB is achieved. The microwave phase quality, such as linewidth, is mainly preserved after conversion. A BER down to \( 10^{-9} \) at 1.25 Gb/s with a detection sensitivity improvement of 9 dB is achieved.

Acknowledgments

S.K. Hwang’s work is supported by the Ministry of Science and Technology of Taiwan under Contract MOST103-2112-M-006-013-MY3 and Contract MOST104-2622-E-006-036-CC2, and also by the Asian Office of Aerospace Research and Development of the U.S. Air Force under Grant FA2386-14-1-0006 and Grant FA2386-15-1-4026.

References

Effects of phase space sticky motions in nearly-integrable dielectric billiards on far-field patterns

Shunya Sekiguchi¹, Susumu Shinohara², Takehiro Fukushima³, and Takahisa Harayama¹

¹Department of Applied Physics, School of Advanced Science and Engineering, Waseda University, 3–4–1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
²NTT Communication Science Laboratories, NTT Corporation, 2–4 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan
³Department of Information and Communication Engineering, Okayama Prefectural University, 111 Kuboki, Soja, Okayama 719-1197, Japan

Email: shunya-s@moegi.waseda.jp, shinohara.susumu@lab.ntt.co.jp, fuku@c.oka-pu.ac.jp, harayama@waseda.jp

Abstract– In the ray dynamics of nearly-integrable billiards, it is well known that survival probability distributions have long time tails due to the sticky motions in the chaotic sea very close to the outermost KAM torus in the phase space. We study how the stickiness influences the emission patterns of nearly-integrable billiard lasers. In this presentation, we will report the relations between the survival probability distributions of ray-chaotic trajectories in the sticky area, the lifetimes and the emission patterns of the resonant modes of the nearly-integrable dielectric billiards.
Tracking of operating point in DC bus system with delayed feedback control for time-varying loads

Koki Yoshida†, Keiji Konishi‡, and Naoyuki Hara†

†Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Naka-ku, Sakai, Osaka 599-8531 JAPAN
Email: konishi@eis.osakafu-u.ac.jp

Abstract—The present paper investigates behavior of DC bus system with delayed feedback control in the case that power consumption of load is changed as a step function and is varied continuously. The delayed feedback controller can track an operating point (i.e., equilibrium point) for a small-amplitude step function, but cannot for a large-amplitude step function. On the other hand, it can track the operating point in the case that the consumption is continuously varied even with large-amplitude. It is shown that, for such time-varying loads, the frequency domain analysis in the field of control theory is useful both to analyze the DC bus system with delayed feedback control and to design the controller.

1. Introduction

Nowadays, considerable attention has been paid to the research of alternating-current (AC) power-grid networks in the field of nonlinear science [1, 2]. In contrast to the AC networks, there has been a growing interest in direct-current (DC) power systems because DC power loads and sources have been widely used in many fields. The DC power systems are now expected to be a future power transmission style [3–5]. Unfortunately, the DC power systems have a serious drawback: the DC bus line voltage may behave oscillatory. This behavior occurs in the following situation: constant power loads (CPLs), which constantly consume DC electric power independent of the line voltage, are connected to the DC bus line [6]. In recent years, CPLs are used everywhere around us, such as IT products. In the field of power electronics, numerous studies have focused on suppression of the oscillations [7–10].

In the past decades, delayed-feedback control [11], one of the most popular methods for stabilizing unstable periodic orbits and unstable equilibrium points embedded within chaotic systems, has received broad attention due to the following advantages: its control law does not require the location of the orbits and the equilibrium points; control signals converge on zero after stabilization [12–14]. Our previous study analyzed dynamics of a simple DC bus system from a viewpoint of bifurcation theory [15]. In addition, our bifurcation analysis indicated that delayed-feedback control can suppress undesired oscillations in bus line by stabilizing an equilibrium point (i.e., an operating point). Our previous study also showed just one numerical example in which delayed-feedback control can track a slow-varying operating point with time-varying consumption in CPLs.

In view of practical situations, power consumption of CPLs must be considered to be varied in time owing to users’ demands. For such practical usage, there is need to know the relationship between the parameters of delayed feedback controllers and the tracking performance. However, our previous study [15] did not provide the relation. The purpose of the present paper is to investigate the relation on numerical simulations. First, we numerically demonstrate that the delayed feedback controller can track an operating point of DC bus system in the case that the power consumption is changed as a step function with small-amplitude. In contrast, with large-amplitude, the controller cannot track it. It is shown that the controller can track it even for large-amplitude change if the consumption is varied slowly. For continuously varying consumption, it is demonstrated that the frequency domain analysis, which is well known as a powerful tool for analyzing the input-output relation of linear systems in the field of control theory, is useful both to analyze the DC bus system with delayed feedback control and to design the controller.

2. DC bus system with a time-varying CPL

Let us consider a DC bus system, as illustrated in Fig. 1, with a delayed feedback controller (i.e., the dotted line rectangle). $E$ represents the DC voltage source, $r$ the equivalent resistance, $L$ the equivalent inductance, $C$ the equivalent capacitance. $v_P(t)$ denotes the bus line voltage, $i_L(t)$ the current through $L$, $i_P(t)$ the current into CPL. The CPL automatically adjusts the current $i_P(t)$ such that the voltage $v_P(t)$ and the current $i_P(t)$ satisfy

$$v_P(t)i_P(t) = P(t), \quad \forall t \geq 0,$$

where $P(t)$ is the power consumed in CPL. Although our previous study treated it as a constant power $P(t) \equiv P$ for simplicity [15], the present study deals with a natural situation, the time-varying power $P(t)$. The delayed feedback controller measures the bus line voltage $v_P(t)$, and then out-
puts the control current,
\[ i_a(t) = \frac{1}{r_k} \{ v_p(t - \Gamma) - v_p(t) \}, \]  
which is proportional to the difference between the present voltage \( v_p(t) \) and the past voltage \( v_p(t - \Gamma) \) with time delay \( \Gamma \geq 0 \).

The circuit equation of the DC bus system with the controller can be described by the dimensionless form
\[
\begin{align*}
\frac{dx}{dt} &= -\frac{a(t)}{x} + by + u, \\
\frac{dy}{dt} &= -x - by + 1,
\end{align*}
\]  
where the control signal is given by
\[ u = k(x_T - x). \]

The state variables and time are described by
\[
\begin{align*}
x &:= \frac{v_p}{E}, \quad y := \frac{L_i}{r_CE}, \quad x_T := \frac{v_p(t - \Gamma)}{E}, \quad u := \frac{ri_u}{E}, \\
\tau &:= \frac{t}{r_CE}, \quad T := \frac{\Gamma}{r_CE}.
\end{align*}
\]  
The system parameters are given as
\[ a(t) := \frac{rP(rC\tau)}{E^2}, \quad b := \frac{r^2C}{L}, \quad k := \frac{r}{r_k}. \]

Consider dynamics of DC bus system (3) without control \((u \equiv 0)\). This system has two equilibrium points,
\[
p_+(\tau) := [x_+^1(\tau), y_+^1(\tau)]^T, \quad p_-(\tau) := [x_-^1(\tau), y_-^1(\tau)]^T,
\]  
\[ x_+^1(\tau) := \frac{1}{2} \left( 1 \pm \sqrt{1 - 4a(\tau)} \right), \quad y_+^1(\tau) := \frac{a(\tau)}{bx_+^1}. \]

It must be noted that locations of the equilibrium points \( p_+(\tau) \) without control do not change even if controller (4) is added to the bus system. However, the local stability of \( p_+(\tau) \) is influenced by controller (4). Our previous study indicated that we have to consider only the stability of \( p_+(\tau) \) owing to instability of \( p_-(\tau) \).

3. Tracking performance for step-type change

This section will show behavior of DC bus system (3) with control in a situation where the parameter \( a(\tau) \) corresponding to the power consumption is changed as a step function,
\[ a(\tau) = \begin{cases} a_L & (\tau < 50) \\ a_H & (\tau \geq 50). \end{cases} \]

Throughout this paper the parameters are fixed to
\[ b = 0.2, \quad k = 0.1, \quad T = 5, \]
and it is supposed that the equilibrium point \( p_+(\tau) \) before jump (i.e., \( a(\tau) = a_L \)) and after jump (i.e., \( a(\tau) = a_H \)) is locally stable. We show two numerical examples below.

First, we have \( a_L = 0.140 \) and \( a_H = 0.185 \). Figure 2(a) shows time-series data of \( a(\tau) \) and \( u(\tau) \). The trajectory of \((x, y)\) is plotted in Fig. 2(b). It can be seen from these figures that \( u(\tau) \) and \((x, y)\) converge on zero and \( p_+(\tau) \), respectively. These results state that controller (4) successfully tracks and stabilizes \( p_+(\tau) \) even if \( a(\tau) \) is changed as a step function.

Second, we change \( a_L = 0.140 \) to \( a_L = 0.120 \). As shown in Fig. 2(b), the trajectory \((x, y)\) does not converge on \( p_+(\tau) \), and then diverges. We have numerically confirmed that controller (4) with parameters (10) successfully tracks and stabilizes \( p_+(\tau) \) only for the range.
$a_{l} \in [0.1292, 0.1973]$. This fact implies that controller (4) fails to track it for $a_{l}$ outside of the range even if $p_{s}(\tau)$ with control at $a_{0}$ is stable. The reason of such unsuccessful tracking would be attributed to the fact that $p_{s}(\tau)$ with control at $a_{0}$ has its own basin of attraction: if $p_{s}(\tau)$ with control at $a_{0}$ is outside of the attraction, controller (4) at $a_{0}$ fails to track it.

Here, instead of the step-type change, $a(\tau)$ is varied slowly from 0.12 to 0.185 with $a(\tau) = 0.1525 - 0.0325 \cos \left[ \pi (\tau - 50)/40 \right]$ for $\tau \in [50, 90]$. Figures 3(a) and 3(b) indicate time series data of $a(\tau)$, $u(\tau)$, $x(\tau)$, and $x^{+}(\tau)$. These figures suggest that controller (4) successfully tracks it for such slow change. From a practical point of view, we should find a critical speed of change below which controller (4) successfully tracks it. In order to find it, we will employ the frequency domain analysis of DC bus system (3) with controller (4) in the next section.

4. Frequency domain analysis

Let us consider the time-varying parameter $a(\tau)$,

$$a(\tau) = a_{0} + a_{s}(\tau),$$

where $a_{0}$ and $a_{s}(\tau)$ denote the nominal value and the external perturbation. The linearized system with control around $p_{s}$ at $a(\tau) = a_{0}$ is described by

$$\frac{d}{d\tau} \begin{bmatrix} x_{\Delta} \\ y_{\Delta} \end{bmatrix} = \begin{bmatrix} -k + a_{0}/(x^{+}_{s})^2 & b \\ -1 & -b \end{bmatrix} \begin{bmatrix} x_{\Delta} \\ y_{\Delta} \end{bmatrix} + \begin{bmatrix} k \\ 0 \end{bmatrix} \begin{bmatrix} x_{\Delta T} \\ y_{\Delta T} \end{bmatrix} + \begin{bmatrix} -1/x^{+}_{s} \\ 0 \end{bmatrix} a_{\Delta s}. \quad (11)$$

where $x_{\Delta} := x - x^{+}_{s}$, $y_{\Delta} := y - y^{+}_{s}$, $x_{\Delta T} := x_{T} - x^{+}_{s}$, $y_{\Delta T} := y_{T} - y^{+}_{s}$. The influence of $a_{\Delta}$ on $x_{\Delta}$ in time-delay linear system (11) can be described by

$$x_{\Delta} = G(s)a_{\Delta}. \quad (12)$$

This transfer function $G(s)$ is given by

$$G(s) := \frac{(s + b)/x^{+}_{s}}{g(s, T)}. \quad (13)$$

The characteristic function $g(s, T)$, which governs the stability of $p_{s}$ at $a(\tau) = a_{0}$, is described by

$$g(s, T) := s^2 + \gamma_{1}s + \gamma_{2} + (\eta_{1}s + \eta_{2})(1 - e^{-sT}), \quad (14)$$

where $\gamma_{1} := b - a_{0}/(x^{+}_{s})^2$, $\gamma_{2} := b - a_{0}b/(x^{+}_{s})^2$, $\eta_{1} := k$, $\eta_{2} := kb$.

Bode diagram of $G(s)$ is plotted in Fig. 4. This diagram provides us plenty of information about tracking performance:

(a) Since we have gain $20\log |G(j\omega)| \leq 6.7$ and phase $\text{Arg}[G(j\omega)] \in (-180^\circ, -154^\circ)$ for low frequency $\omega \leq 0.10$, then amplitude and phase of $x$ are almost same as those of $x^{+}_{s}$ when $a(\tau)$ is varied with $\omega \leq 0.1$.

(b) Since we have the peak gain $20\log |G(j\omega)| = 23.3$ at the resonant frequency $\omega = 0.35$, the amplitude of $x$ becomes large when $a(\tau)$ is varied around the resonant frequency.

(c) Since we have phase $\text{Arg}[G(j\omega)] = -270^\circ$ above the resonant frequency $\omega = 0.35$, the phase of $x$ is different from that of $x^{+}_{s}$ when $a(\tau)$ is varied with $\omega \geq 0.35$.

The information provided above is verified by some numerical examples below.

The parameter $a(\tau)$ is varied as

$$a_{0} = 0.17, \ a_{s}(\tau) = 0.015 \sin \{\omega(\tau - 50)\}.$$
Figure 5: Time series data of $x$ and $x^+$ for periodic $a(\tau)$ with (a) $\omega = 0.05$, (b) $\omega = 0.25$, and (c) $\omega = 0.70$.

Figure 5(a) shows time series data of $x$ and $x^+$ for periodic $a(\tau)$ with $\omega = 0.05 < 0.10$. It can be seen that the amplitude and phase of $x$ are almost same as those of $x^+$. This numerical result agrees well with information (a) on the bode diagram. The time series data near the resonant frequency, $\omega = 0.25$, are shown in Fig. 5(b). We see the amplitude of $x$ is large. This numerical result also agrees with information (b). Figure 5(c) shows time series data above the resonant frequency, $\omega = 0.70 > 0.35$. The phase of $x$ is different from that of $x^+$; information (c) agrees with our time series data.

The information on bode diagram can be used in two ways: we can estimate the upper limit of perturbation frequency of $a(\tau)$ for given controller parameters $(k, T)$; on the other hand, we can design $(k, T)$ such that a given perturbation frequency of $a(\tau)$ is less than the resonant frequency. It can be stated that the information on bode diagram is useful for design of such tracking controller.

5. Conclusion

We conclude that the delayed feedback controller can track the operating point of DC bus system in the case that the power consumption is changed as a small step function and is slowly varied with large-amplitude. It has been shown that the frequency domain analysis provides us plenty of information about tracking performance. The information was verified by some numerical examples.

Acknowledgments

This research was partially supported by JSPS KAKENHI (26289131).

References

Stability analysis of amplitude death in Cartesian product networks of delay-coupled Landau-Stuart oscillators

Yoshiki Sugitani

Department of Electrical and Electronic Engineering, Ibaraki University
4–12–1 Nakanarusawa, Hitachi, Ibaraki 316-8511 Japan
Email: yoshiki.sugitani.0301@vc.ibaraki.ac.jp

Abstract—This report investigates amplitude death in Cartesian product networks of delay-coupled oscillators. The Cartesian product networks consist of two sub-networks which have different connection delays with each other; that is, the connection delays are not identical in the whole networks. Although such networks are difficult to analyze, the feature of the Cartesian product networks allows us to analyze such networks easily. The analytical result is confirmed by numerical simulations.

1. Introduction

The various collective phenomena in coupled oscillators have been widely investigated in biological, physical, chemical, and social systems [1]. One of such phenomena is amplitude death where a homogeneous steady state in coupled oscillators is stabilized by diffusive connections; that is, the oscillations of all the oscillators are quenched. Although amplitude death never occurs in diffusively-coupled identical oscillators [2], connection delays can cause amplitude death even in coupled identical oscillators [3].

Amplitude death induced by the connection delays has been great interest in nonlinear science [4]. Various types of delay connections that cause amplitude death have been proposed such as the distributed-delay connection [5, 6], the multiple-delay connection [7], the time-varying delay connection [8], the integrated delay connection [9], the multicomponent delay connection [10], the integrated delay connection [11], and the mixed time-delay connection [12]. All the previous studies assumes that all the connection delays are identical in the whole network. In the real world, however, it is totally impractical that all the connection delays are identical. Generally, it is difficult to analyze the coupled oscillators with non-identical connection delays.

The Cartesian product is one of the basic operation on Graph theory [13]. By using Cartesian product, we can construct various complex networks from simpler sub-networks, for instance, regular grids are constructed from two path graphs. It is well-known that the eigenvalues of the Laplacian matrix of a Cartesian product network are calculated by the sum of the eigenvalues of its sub-networks. Based on this fact, some researchers have investigated partial and full synchronization in Cartesian product networks of coupled oscillators [14, 15, 16].

In this report, we investigate amplitude death in Cartesian product networks of delay-coupled oscillators. The Cartesian product networks consist of two sub-networks which have different connection delays with each other. Therefore, the connection delays are not identical in the whole networks. Even in such situation, the feature of Cartesian product allows us to easily analyze the stability of amplitude death. Furthermore, it is shown that the stability of amplitude death is heavily depends on the topology of the sub-networks.

The following notations are used throughout this report. \( \mathcal{G} = (\mathcal{V}, \mathcal{E}) \) is the graph consisting of set of nodes \( \mathcal{V} \) and edges \( \mathcal{E} \). Conversely, \( \mathcal{V}(\mathcal{G}) \) and \( \mathcal{E}(\mathcal{G}) \) represent the sets of nodes and edges of the graph \( \mathcal{G} \), respectively. \( A_{\mathcal{G}} \) is the adjacency matrix of graph \( \mathcal{G} \): if \( i \)-th and \( l \)-th nodes are connected by an edge, then \( [A_{\mathcal{G}}]_{il} = [A_{\mathcal{G}}]_{li} = 1 \); otherwise, \( [A_{\mathcal{G}}]_{il} = [A_{\mathcal{G}}]_{li} = 0 \). The matrix \( I_N \) denotes the \( N \times N \) unit matrix. The imaginary unit is defined as \( j := \sqrt{-1} \).

2. Cartesian Product network of delay coupled oscillators

This section briefly introduces the Cartesian product network consisting of two sub-networks. Then, we will explain the delayed coupled oscillators of the Cartesian product network where the two sub-networks have different connection delays.
2.1. Cartesian product network

The Cartesian product network consisting of two sub-networks \( G_1 = (V(G_1), E(G_1)) \) and \( G_2 = (V(G_2), E(G_2)) \) is denoted by \( G_1 \square G_2 \). The nodes set of \( G_1 \square G_2 \) is given by \( V(G_1 \square G_2) = V(G_1) \times V(G_2) \). Edge \((v_1, v_2), (v_1', v_2')\) is an edge in \( E(G_1 \square G_2) \) if \( v_1 = v_1' \) and \((v_2, v_2') \in E(G_2) \) (or if \( v_2 = v_2' \) and \((v_1, v_1') \in E(G_1) \)). Figure 1 shows an example of the Cartesian product network. The adjacency matrix of the Cartesian product network \( G_1 \square G_2 \) is given by

\[
A_{G_1 \square G_2} = A_{G_1} \otimes I_n + I_m \otimes A_{G_2},
\]

where \( m \) and \( n \) are the number of nodes in sub-networks \( G_1 \) and \( G_2 \), respectively. The symbol \( \otimes \) denotes the Kronecker product.

2.2. Delayed coupled oscillators

Let us consider the delayed coupled oscillators of the Cartesian product network \( G_1 \square G_2 \) illustrated in Fig. 1. The dynamics of the oscillators are given by:

\[
\dot{Z}_i(t) = (a + j\omega - |Z_i(t)|^2)Z_i(t) + u_i^{(1)}(t) + u_i^{(2)}(t),
\]

where \( Z_i(t) \in \mathbb{C} \) is the state variables of \( i \)-th oscillators. \( a > 0 \) and \( \omega > 0 \) respectively represent instability of the fixed point \( Z_i^* = 0 \) and the natural frequency of oscillators. Each oscillator receives the input signal \( u_i^{(1)}(t) \) and \( u_i^{(2)}(t) \) from sub-networks \( G_1 \) and \( G_2 \), respectively.

\[
u_i^{(1)}(t) = k \left( \frac{1}{d_i^{(1)}} \sum_{l=1}^{m} c_{i,l}^{(1)} Z_l(t - \tau_1) \right) - Z_i(t),
\]

\[
u_i^{(2)}(t) = k \left( \frac{1}{d_i^{(2)}} \sum_{l=1}^{m} c_{i,l}^{(2)} Z_l(t - \tau_2) \right) - Z_i(t),
\]

where \( k \) is the coupling strength. \( \tau_1 \) and \( \tau_2 \) denote the connection delays in sub-networks \( G_1 \) and \( G_2 \), respectively (see Fig. 1). Note that the connection delays would differ from sub-network to sub-network. \( c_{i,l}^{(1)} \) and \( c_{i,l}^{(2)} \) are \((i,l)\) elements of adjacency matrix \( A_{G_1} \otimes I_n \) and \( I_m \otimes A_{G_2} \), respectively. \( d_i^{(1,2)} \) represent the degree of \( i \)-th oscillator in sub-networks \( G_1 \) and \( G_2 \). The coupled oscillators (2), (3) have the homogeneous steady state

\[
[Z_1^*, \ldots, Z_m^*]^T = [0, \ldots, 0]^T.
\]

3. Linear stability analysis

Linearizing Eqs. (2) and (3) around steady state (4), we obtain

\[
\dot{z}_i(t) = \left( a + j\omega - 2k \right) z_i(t) + \frac{k}{d_i^{(1)}} \sum_{l=1}^{m} c_{i,l}^{(1)} z_l(t - \tau_1) + \frac{k}{d_i^{(2)}} \sum_{l=1}^{m} c_{i,l}^{(2)} z_l(t - \tau_2),
\]

where \( z_i(t) = Z_i(t) - Z_i^* \) is the perturbation from steady state (4). Linear system (5) can be rewritten as

\[
\dot{X}(t) = (a + j\omega - 2k)X(t) + k \left( E_1 \otimes I_n \right) X(t - \tau_1) + k \left( I_m \otimes E_2 \right) X(t - \tau_2),
\]

where \( X(t) := [z_1(t), \ldots, z_m(t)]^T \). The matrices \( E_1 := D_1^{-1} A_{G_1} \) and \( E_2 := D_2^{-1} A_{G_2} \) denote the network topologies of sub-networks \( G_1 \) and \( G_2 \), where \( D_1, D_2 \in \mathbb{R}^{mn \times mn} \) is the diagonal matrix of oscillator's degree, i.e., its \( i \)-th diagonal element is the degree of \( i \)-th oscillator on sub-network \( G_1 (G_2) \).

The stability of linear system (6) is governed by the roots of the following characteristic equation.

\[
G(s) := \text{det} \left[ \begin{pmatrix} (s - a - j\omega + 2k)I_{nm} - \left( E_1 \otimes I_n \right) & \left( I_m \otimes E_2 \right) e^{-s\tau_2} \end{pmatrix} \right].
\]

It is known that the matrices \( E_1 \) and \( E_2 \) can be diagonalized as follows [17],

\[
T_1^{-1} E_1 T_1 = \text{diag}(\rho_1, \ldots, \rho_m),
\]

\[
T_2^{-1} E_2 T_2 = \text{diag}(\sigma_1, \ldots, \sigma_n),
\]

where \( T_1 \) and \( T_2 \) are transformation matrices. \( \rho_1, \ldots, \rho_m \) and \( \sigma_1, \ldots, \sigma_n \) denote the eigenvalues of \( E_1 \) and \( E_2 \), respectively. The matrices \( E_1 \) and \( E_2 \) in Eq. (7) can be simultaneously diagonalized by using the transformation matrix \( T_1 \otimes T_2 \) as follows:

\[
G(s) = \text{det} \left[ \begin{pmatrix} T_1^{-1} \otimes T_2^{-1} \end{pmatrix} \begin{pmatrix} (s - a - j\omega + 2k)I_{nm} - \left( \text{diag}(\rho_1, \ldots, \rho_m) \otimes I_n \right) e^{-s\tau_1} + (I_m \otimes E_2) e^{-s\tau_2} \end{pmatrix} \right]
\]

\[
= \text{det} \left[ \begin{pmatrix} (s - a - j\omega + 2k)I_{nm} - \left( \text{diag}(\rho_1, \ldots, \rho_m) \otimes I_n \right) e^{-s\tau_1} + (I_m \otimes \text{diag}(\sigma_1, \ldots, \sigma_n)) e^{-s\tau_2} \end{pmatrix} \right].
\]

This diagonalization allows us to separate the characteristic Eq. (7) into \( mn \) modes,

\[
G(s) = \prod_{p=1}^{m} \prod_{q=1}^{n} g(s, \rho_p, \sigma_q),
\]

where

\[
g(s, \rho, \sigma) := (s - a - j\omega + 2k - k(\rho e^{\tau_1}) + k(\sigma e^{\tau_2})).
\]

As a consequence, steady state (4) is stable if and only if all the \( mn \) modes of Eq. (8) is stable.

For checking the stability of Eq. (8), we will focus on roots of \( g(s, \rho, \sigma) \neq 0 \). The stability of Eq. (8) changes only when the roots crosses the imaginary axis. Substituting \( s =
$\tau_1$ and $\tau_2$ into $g(s, \rho, \sigma) = 0$ gives us the following two equations,

\begin{align*}
-a + 2k - k\rho \cos(\tau_1) - k\sigma \cos(\tau_2) &= 0, \\
\lambda - \omega + k\rho \sin(\tau_1) + k\sigma \sin(\tau_2) &= 0.
\end{align*}

(9)

Solving Eq. (9) in terms of $\tau_1$ and $\tau_2$ yields the marginal stability curves on the connection parameter $(\tau_1, \tau_2)$ space [7]. Moreover, in order to derive the stability region from the marginal stability curves, we have to check the direction of the roots crossing the imaginary axis. The direction can be checked by the real part of $\text{d}s/\text{d}\tau_2$ at $s = j\lambda$,

$$\text{Re}\left[\frac{\text{d}s}{\text{d}\tau_2}\right]_{s=j\lambda} = \text{Re}\left[-\frac{j\lambda k e^{j\tau_2}}{1 + k(\rho \tau_1 e^{j\tau_1} + \sigma \tau_2 e^{-j\tau_2})}\right].$$

(10)

The positive (negative) sign of Eq. (10) denotes that the roots crossing the imaginary axis from left to right (right to left) as $\tau_2$ increases.

4. Numerical examples

For numerical examples, we consider two Cartesian product networks A and B illustrated in Fig. 2. Both of the networks have the same number of oscillators and same sub-network $G_2$; that is, they have different sub-network $G_1$. The eigenvalues of $E_1$ for Network A are $\rho_1, 2 = -0.5, \rho_3 = 1$. Those for Network B are $\rho_1 = -1, \rho_2 = 0, \rho_3 = 1$. The eigenvalues of $E_2$ are $\sigma_1 = -1, \sigma_2 = 1$. Throughout this report, the parameters of oscillators (2) and the coupling strength are fixed at

$$a = 0.50, \quad \omega = \pi, \quad k = 2.0.$$

(11)

Figure 3 shows the marginal stability curves and the stability region on the connection parameter $(\tau_1, \tau_2)$ space. The thin (bold) curves denote that when a parameter set $(\tau_1, \tau_2)$ crosses the curves with increasing $\tau_2$, one root of $g(s, \rho, \sigma) = 0$ crosses the imaginary axis from left to right (right to left). The shaded area shows the stability region where all the roots of $G(s)$ are located on the left-half of the complex plane. In other words, the local stability of steady state (4) is guaranteed in this region.

Comparing Fig. 3(a) with Fig. 3(b), the region for Network B is symmetry about the slanted line $\tau_1 = \tau_2$, while it is not symmetry for Network A. Moreover, the region for Network A has the range of $\tau_1$, which is between two white dotted lines in Fig 3(a), such that we can use the long connection delay $\tau_2$ of sub-network $G_2$ to induce amplitude death. It should be noted that long connection delays never induce amplitude death if the connection delays are identical (i.e., $\tau_1 = \tau_2$) in the whole networks [3].

Figure 4 shows the time-series data of the state variables $\text{Re}[Z(t)]$ at points (a):$(\tau_1, \tau_2) = (2.0, 2.0)$ and (b):$(\tau_1, \tau_2) = (0.75, 2.0)$ in Fig. 3(a). At $t = 30$, all the oscillators are coupled. For point (a), the variables still oscillate after coupling. For point (b), they converge onto steady state (4).

\textsuperscript{1}We can use even a diffusive connection (i.e., $\tau_2 = 0$).
5. Conclusion

This report has investigated amplitude death in Cartesian product networks of delayed coupled oscillators, where two sub-networks of the Cartesian product networks have different connection delays with each other. By using the feature of Cartesian product, we have easily analyzed the local stability of the steady state. The analytical results were numerically confirmed.

References

Stability of Paralleled Boost Converters with WTA Switching

Yasuo Murata† and Toshimichi Saito †
†Department of Electronics and Electrical Engineering, Hosei University
3–7–2 Kajino-cho, Koganei, Tokyo 184-8584, Japan
Emails: yasuo.murata.6c@stu.hosei.ac.jp; tsaito@hosei.ac.jp

Abstract—This paper studies a paralleled system of boost converters with WTA-based switching rule. The system exhibits multi-phase synchronization phenomena and chaotic phenomena. The multi-phase synchronization is suitable for ripple reduction, current sharing, and efficient power supply. The WTA-based switching rule is effective to reinforce the fault tolerance. Simplifying the system into a piecewise linear model, stability of the synchronization phenomena and ripple waveforms can be analyzed precisely.

1. Introduction

The paralleled systems of switching power converters have been studied from fundamental and application viewpoints. In the fundamental study, the paralleled systems are interesting examples of switched dynamical systems that can exhibit a variety of nonlinear phenomena such as multi-phase synchronization and chaos [1]-[4]. In the applications, the paralleled systems can realize current sharing and ripple reduction which are effective in robust and reliable power management [5]-[9]. In these studies, analysis of nonlinear phenomena is important and single power converters have been studied sufficiently. However, the analysis of the paralleled systems is not easy because they are higher dimensional nonlinear systems with various complex behavior.

This paper studies stability of a paralleled system of N boost converters through which N pieces of input voltage sources are applied to a load. The N boost converters are coupled by the winner-take-all (WTA) switching rule that can realize N-phase synchronization (N-SYN) automatically. The N-SYN is suitable for ripple reduction of input current that is effective to realize higher efficiency operation. Especially, if the input voltage is given by solar cells, the rippled reduction is well suited for maximum power point tracking [10] [11]. The parallel operation is also suitable in viewpoint of fault-tolerance: if some converter is broken, the other converters can preserve the operation. In order to analyze the system, we introduce a simple piecewise linear model and provide a sufficient condition of parameters for stability of N-SYN and precise calculation formula of the input ripple. These results can clarify stable operation, power efficiency and fault tolerance. Performing basic numerical experiments, the stability of N-SYN and robust operation of the circuit are confirmed. We have prepared laboratory measurements of typical phenomena for the final version.

2. 3-Paralleled Boost Converter

Fig. 1 shows the simplified circuit model of the paralleled boost converters where \( r_L \) is an inner resistance of inductor. The \( j \)-th converter includes the switch \( S_j \) and diode \( D_j \) which can be either of the State A or State B: where \( j = 1 \sim N \).

State A: \( S \) conducting, \( D \) blocking, and \( i_j > 0 \).
State B: \( S \) blocking, \( D \) conducting, and \( i_j > 0 \). (1)

For simplicity, let \( N = 3 \) hereafter. Let \( i_{in} = i_1 + i_2 + i_3 \) be an input current. In order to simplify the analysis, \( RC \) load is replaced with constant voltage sources \( V_{oj} \) shown in Fig. 3 where \( RC \gg T \) is assumed. \( T \) is a clock period.

This circuit dynamics is described by

\[
\frac{Ldi_j}{dt} = \begin{cases} 
-ri_j + V_{in} & \text{for State A} \\
-ri_j - V_{oj} & \text{for State B} 
\end{cases}
\] (2)

Eq. (6) is changed into the following dimensionless equation because the analysis is simplified.

\[
\frac{dx_j}{d\tau} = \begin{cases} 
-\gamma x_j + a & \text{for State A} \\
-\gamma x_j - b_j & \text{for State B} 
\end{cases}
\] (3)
Where the following dimensionless variables and parameters are used:

\[ a = \frac{TV_{in}}{LJ}, \quad b_j = \frac{T(V_{o1} - V_{in})}{LJ}, \quad \gamma_j = \frac{T_{rL}}{L} \]  
\[ \tau = \frac{i}{T}, \quad x_j = \frac{i_j}{J}, \quad X_n = \frac{J}{T} \]

Let \( x_j = \frac{i_{in}}{J} = x_1 + x_2 + x_3 \) be a dimensionless input current. In the circuit, \( J > 0 \) is the lower current threshold and \( J > 0 \) is a current criterion for a desired operation. The state transition is defined by following switching rule. In the State A, the inductor current \( i_j \) rises as shown in Fig. 3. If the \( X_n \) is the maximum among \( t \) to \( x_N \) at some clock signal arriving time \( \tau = n \) then State A is changed into State B. Let the \( j \)-th system be State B where \( j \)-th dimensionless current \( x_j \) decays. If the \( x_j \) reaches the lower threshold \( X_n \) then the State B changed into State A. Note that the three converters are connected by the comparison of \( x_1 \sim x_3 \) (\( i_1 \sim i_3 \)) in the SW rule.

**SW rule**

\[
\text{State A} \rightarrow \text{State B if } i_j = \text{Max} (t = nT) \\
\text{State B} \rightarrow \text{State A if } i_j = J.
\]

The piecewise exact solution is given by

\[
x_j(\tau) = \begin{cases} 
(x(0) - a/\gamma)e^{-\gamma \tau} + a/\gamma & \text{for State A} \\
(x(0) + b_j/\gamma)e^{-\gamma \tau} - b_j/\gamma & \text{for State B}
\end{cases}
\]

If \( Df_0(a, b) \) is larger than 1, the N-SYN is unstable where \( x(0) \) indicates an initial value. Using these equations, we can calculate waveforms precisely. In this paper, the parameters condition is following.

\( a = 0.3, \quad \gamma = 0.3, \quad X_n = 0.05, \quad b_j : \text{varies} \)

Fig. 4 shows typical waveforms of 3-phase synchronization phenomenon.

**3. Stability of N-phase synchronization**

Here we define the N-phase synchronization (N-SYN) for \( N = 3 \). Let \( x = (x_1, \ldots, x_N) \). \( x \) is said to be N-SYN if

\[
x(\tau + 3) = x(\tau)
\]

\[
x_2(\tau) = x_1(\tau + 1), \quad x_3(\tau) = x_2(\tau + 1) \quad \text{or}
\]

\[ x_3(\tau) = x_1(\tau + 1), \quad x_2(\tau) = x_1(\tau + 1) \]

Conditions to be N-SYN is \( b_1 = b_2 = b_3 = b \). The N-SYN is stable if

\[
Df_0(a, b) \equiv \left| \frac{X_1 - P_1}{X_n + P_2}e^{-aX_n} \right| < 1
\]

where \( P_1 = a/\gamma > 0, \quad P_2 = b_j/\gamma > 0, \quad 0 < X_n < P_1 \).

Fig. 5 and 6 show the stable factor \( Df \) for \( N = 3 \) and \( N = 2 \), respectively. Fig. 7 (a) shows a waveform of 3-SYN with input current. In Figs. 5 and 6, the bifurcation diagram is given by peak of \( x_1 \), indicated by \( x_n \) in Fig. 7 (a). In Fig. 5, 3-SYN is stable (unstable) for \( b > b_p \) (\( b < b_p \)) where \( Df_1(0.3, b_p) = 1 \). In Fig. 6, 2-SYN is stable (unstable) for \( b > b_q \) (\( b < b_q \)) where \( Df_2(0.3, b_q) = 1 \). Note that \( b_p < b_q \).
That is, both 3-SYN and 2-SYN are stable for \( b > b_q \). 3-SYN is stable and 2-SYN is unstable for \( b_q > b > b_p \). Both 3-SYN and 2-SYN are unstable for \( b_p > b \).

Here we assume that an accident occurs and the third converter is broken in the case of where 3-SYN is stable. The system of 3 converters is changed into that of 2 converters. If \( b > b_q \) then 2-SYN is stable after the accident as shown in Fig. 7 (a') where ripple is reduced after the accident. If \( b_q > b > b_p \) then 2-SYN is unstable after the accident as shown in Fig. 7 (b'). If \( b_p > b \) then unstable 3-SYN is changed into hyperchaos as shown in Fig. 7 (c').

In order to consider the dynamics, we introduce the Lissajous figure as shown in Fig. 8. The Lorenz plot suggests complicated dynamics. The failed converter can exhibit interesting chaotic phenomena as shown in Fig. 7 (c') and Fig. 8 (c).

Figure 5: Bifurcation diagram from/to 3-SYN. \((a = 0.3, \gamma = 0.3)\)

Figure 6: Bifurcation diagram from/to 2-SYN. \((a = 0.3, \gamma = 0.3)\)

Figure 7: Typical waveforms \((a = 0.3, \gamma = 0.3)\) (a) Stable 3-SYN and input current \( b = 0.20 \) before the accident (a'). Stable 2-SYN \( b = 0.20 \) after the accident. (b) Stable 3-SYN \( b = 0.12 \) (b') Unstable 2-SYN \( b = 0.12 \). (c) Unstable 3-SYN \( b = 0.045 \) (c') Hyperchaos \( b = 0.045 \).
4. Conclusion

Introducing a simple PWL model of paralleled boost converters. Stability of N-SYN and fault-tolerance have been considered in this paper. It is confirmed that the WTA-switching rule can preserve N-SYN after the accident in some parameter a range. Future problems includes more detailed analysis of stability, ripple characteristic, and power efficiency. Now we are preparing laboratory experiments for confirmation of typical phenomena.
Basins of Attraction of Steady Operating Conditions in a Two-site Electricity and Heat Supply System

Hikaru Hoshino† Yoshihiko Susuki‡§ and Takashi Hikihara†

†Department of Electrical Engineering, Kyoto University
Katsura, Nishikyo, Kyoto 615-8510 Japan
‡Department of Electrical and Information Systems, Osaka Prefecture University
1-1 Gakuencho, Nakaku, Sakai, Osaka, 599-8513 Japan
§JST-CREST 4-1-8 Honcho, Kawaguchi, Saitama 332-0012 Japan

Email: hoshino@dove.kuee.kyoto-u.ac.jp, susuki@eis.osakafu-u.ac.jp, hikihara.takashi.2n@kyoto-u.ac.jp

Abstract—This paper analyzes the basin of attraction of a stable equilibrium point representing a steady operating condition of synchronous generators in a two-site electricity and heat supply system. The analysis is used for considering the effect of heat transfer management on the dynamics of the generators. The basin of attraction becomes small depending on the heat transfer rate, and a change of the set-points of the combined heat and power plants for regulating the heat transfer rate possibly destabilizes the generators.

1. Introduction

This paper numerically studies a dynamical model of two-site electricity and heat supply system based on our previous studies [1–3]. We examine the basin of attraction of a stable equilibrium point representing a steady operating condition of synchronous generators in the electric sub-system. This is of basic significance for understanding the system’s response to an open-loop control of the energy flows in the two-site system. In [3], we proposed a state-feedback (closed-loop) controller that enables regulation of electricity and heat flows based on the nonlinear control technique [4, 5]. This controller provides a trajectory of state variable which realizes the desired energy flows. However, it seems not easy to estimate the basin of attraction of the desired trajectory because of the complexity of the closed-loop system, in which the dynamics of the electric and heat sub-systems are coupled each other.

In this paper, we consider an open-loop control in which the set-points of Combined Heat and Power (CHP) plants (including gas turbines and generators) are already determined to realize the desired energy flows. To the open-loop control, the responses of the electric and heat sub-systems can be considered separately. In a viewpoint of dynamical systems theory, the studied model of the electric sub-system appears as a model of double swing dynamics with external forcing [6–8]. In [6–8], the basin structure of the model of swing dynamics was investigated by taking systematic slices of the phase space. In [9], a similar method for analyzing basin structure of dynamical systems is developed by using cell state space and mapping on it. Based on these studies, in [2], the basin portraits of the dynamical model of the two-site system was visualized under several fixed values of the set-points of the CHP plants. The visualization was then used for understanding an ideal response of the two-site system to a step-wise change of the set-points of the CHP plants. Here, we discuss this in a more realistic situation, and consider a ramp-wise change of the set-points.

2. Mathematical model

This section introduces a dynamical model of the electricity and heat supply system based on [1, 2]. Figure 1 shows the block diagram of the two-site system in which the positive directions of energy flows are denoted. The model of the electric sub-system was visualized under several fixed values of the set-points of the CHP plants. The visualization was then used for understanding an ideal response of the two-site system to a step-wise change of the set-points of the CHP plants. Here, we discuss this in a more realistic situation, and consider a ramp-wise change of the set-points.

2.1. Electric sub-system

The electric sub-system in Fig. 1 consists of the two generators, power loads, transmission lines, and infinite bus. The model of electric sub-system is based on the swing equation [10] with \( \delta_i \) representing the electric angular position of rotor with respect to the infinite bus, and \( \omega_i \) the
deviation of rotor speed relative to the synchronous speed \( \omega_s \). The variable \( \delta_i \) is in the electrical radian, and \( \omega_i \) is scaled by \( \omega_i := \sqrt{\omega_i/2H_i} \), where \( H_i \) stands for the per-unit time constant of rotor. The dynamics of generators are represented as follows: for \( i = 1, 2 \),

\[
\frac{d\delta_i}{dt} = \omega_i, \quad \frac{d\omega_i}{dt} = P_{mi} - D_i \omega_i - P_{e_i}(\delta_1, \delta_2), \tag{1}
\]

where \( P_{mi} \) stands for the mechanical input power to the generator, and \( D_i \) for the damping coefficient. The function \( P_{e_i} \) stands for the electric output power and is given by

\[
P_{e_i} = \sum_{j \in \{1, 2, \infty\}} E_j E_j(G_{ij} \cos(\delta_i - \delta_j) + B_{ij} \sin(\delta_i - \delta_j)), \tag{2}
\]

with the symbol \( \infty \) representing the infinite bus, and \( \delta_{\infty} = 0 \). The parameter \( E_i \) corresponds to the voltage behind synchronous reactance, and \( G_{ij} + iB_{ij} \) are the transfer admittances.

2.2. Heat sub-system

The heat sub-system in Fig. 1 consists of the condution pipe and loads. Here, we do not consider the transient dynamics and losses of heat transfer through the heat conduction pipe. This is relevant for considering the open-loop control of the two-site system. By using the following model, the set-points, i.e. the fuel inputs to the CHP plants, are determined to realize a desired heat transfer rate \( Q'_e \). In Fig. 1, the conservation of energy at each site induces the following equality:

\[
Q'_{chp1} = Q'_e + Q'_{L1}. \tag{3}
\]

Further, the heat output rates \( Q'_{e1} \) and \( Q'_{e2} \) satisfy

\[
Q'_{e1} = -Q'_{e2} := Q'_e, \tag{4}
\]

where \( Q'_e \) represents the heat transfer rate from site #1 to site #2.

2.3. Gas turbine

The gas turbine at site \( i \) converts the gas input rate \( P_{gasi} \) to both the mechanical power \( P_{mi} \) and the heat rate \( Q'_{chp} \). Because its time response is sufficiently fast compared with the electromechanical dynamics of the generators [11], the dynamics of the gas turbine are not considered in this paper. Then, the instantaneous conversion of energy at each gas turbine is represented by

\[
\begin{bmatrix}
P_{mi} \\
Q'_{chp}
\end{bmatrix} =
\begin{bmatrix}
\eta_{ei} \\
\eta_{hi}
\end{bmatrix} P_{gasi}.
\tag{5}
\]

Throughout this paper, the parameters \( \eta_{ei} \) and \( \eta_{hi} \) are constant and satisfy \( \eta_{ei} + \eta_{hi} < 1 \). The constant \( \eta_{ei} \) represents the thermal efficiency of the gas turbine at site \( i \), and \( \eta_{hi} \) the ratio of heat output rate to gas input rate.

2.4. Derived model

Consequently, the dynamics of the two-site electricity and heat supply system are represented by the following nonlinear dynamical model:

\[
\begin{align*}
\frac{d\delta_1}{dt} &= \omega_1, \quad \frac{d\omega_1}{dt} = \frac{\eta_{ei}}{\eta_{hi}} (Q'_e + Q'_{L1}) - D_1 \omega_1 - P_{e1}(\delta_1, \delta_2), \tag{6a}
\frac{d\omega_2}{dt} &= \omega_2, \quad \frac{d\delta_2}{dt} = \frac{\eta_{ei}}{\eta_{hi}} (Q'_e + Q'_{L2}) - D_2 \omega_2 - P_{e2}(\delta_1, \delta_2). \tag{6b}
\end{align*}
\]

The dynamical model (6) contains the parameters \( Q'_e \) and \( Q'_{L1} \) of the heat sub-system. In the rest of this paper, with this model, the effect of the heat sub-system on dynamics of the electric sub-system will be studied.

3. Steady operating conditions

This section analyzes equilibrium points of the dynamical model (6) in order to investigate the steady operating conditions of the generators. Since the dynamical model (6) has the same formulation as the classical swing equations, the analysis method used in [12] is applied for investigating how the steady state characteristics depend on \( Q'_e \). From the condition \( d\delta_i/dt = 0 \) at equilibrium points, we have

\[
\omega_i^* = 0, \tag{7}
\]

where \( \omega_i^* \) represents the value of \( \omega_i \) at equilibrium points. From the condition \( d\omega_i/dt = 0 \), the values of phase angles \( \delta_i^* \) and \( \delta_j^* \) satisfy the following equations:

\[
\begin{align*}
\alpha_1 &= \sin \delta_1^* + \kappa_1 \sin(\delta_1^* - \delta_2^*) + \lambda_1 \cos \delta_1^* + \mu_1 \cos(\delta_1^* - \delta_2^*), \\
\alpha_2 &= \sin \delta_2^* + \kappa_2 \sin(\delta_2^* - \delta_1^*) + \lambda_2 \cos \delta_2^* + \mu_2 \cos(\delta_2^* - \delta_1^*),
\end{align*}
\tag{8}
\]

where \( \alpha_1 \) and \( \alpha_2 \) are defined by

\[
\begin{align*}
\alpha_1 &= \frac{\eta_{ei}(Q'_{L1} + Q'_e) - \eta_{hi} E_1^2 G_{11}}{\eta_{hi} E_1 E_2 B_{110}}, \tag{9a}
\alpha_2 &= \frac{\eta_{ei}(Q'_{L2} - Q'_e) - \eta_{hi} E_2^2 G_{22}}{\eta_{hi} E_2 E_3 B_{210}}, \tag{9b}
\end{align*}
\]

and \( \kappa_1, \lambda_1, \mu_1 \) and \( \kappa_2, \lambda_2, \mu_2 \) are given by

\[
\kappa_i = \frac{E_1 E_2 B_{12}}{E_1 E_3 B_{110}}, \quad \lambda_i = \frac{G_{11}}{B_{110}}, \quad \mu_i = \frac{E_1 E_2 G_{12}}{E_1 E_3 B_{110}}. \tag{10}
\]

By solving the equation (8), the values of \( \delta_1^* \) and \( \delta_1^* \) are numerically determined. Fig. 2 shows the result on existence and number of equilibrium points. The values of parameters are shown in Tab. 1. In the region \( \mathbb{R}_n \) (\( n = 2, 4, 6 \), there are \( n \) distinct equilibrium points. In the three regions, one of the equilibrium points is asymptotically stable, and the others are unstable. The stable equilibrium point represents a synchronized motion of the two generators in which they operate with the same frequency as the infinite bus.

- 676 -
Here, we consider the stability of the equilibrium points due to the quasi-static changes of the parameters of the heat sub-system. As $Q'_n$ changes, the steady operating point moves in the $(\alpha_1, \alpha_2)$-plane along the straight line given by

$$e_1 \alpha_1 + e_2 \alpha_2 = (Q'_{L,1} + Q'_{L,2}) - e_3$$

(11)

where the coefficients $e_1$ to $e_3$ are determined by the parameters of the electric sub-system and are given by

$$e_1 := \frac{\eta_1}{\eta_{L,1}} E_1 E_{\infty} B_{1,\infty}, \quad e_2 := \frac{\eta_2}{\eta_{L,2}} E_2 E_{\infty} B_{2,\infty},$$

$$e_3 := \frac{\eta_1}{\eta_{L,1}} E_1^2 G_{11} + \frac{\eta_2}{\eta_{L,2}} E_2^2 G_{22}.$$  
(12)

The equation (11) is obtained by eliminating $Q'_n$ from (9). Since the line (11) is parameterized by $Q'_{L,1} + Q'_{L,2}$, a steady operating condition is determined by the values of $Q'_n$ and $Q'_{\text{sum}} := Q'_{L,1} + Q'_{L,2}$. In Fig. 2, the red line shows (11) with $Q'_{\text{sum}} = 1.8$. The synchronized operation of the generators is achieved when the operating condition determined by $Q'_n$ and $Q'_{\text{sum}}$ is kept within $R_2$.

Here, we consider the stability of the equilibrium points under several fixed values of $Q'_n$. Under the current setting of the parameters, the system (6) has four attractors. One attractor is the stable equilibrium point representing the steady operating condition: this is shown by circle (○) in the figure, and its basin is colored green. A second attractor is a periodic orbit, in which the generator #1 operates at a desynchronized manner with the infinite bus: its basin is colored red. A third one is another periodic orbit, in which the generator #2 is desynchronized; its basin is colored orange. In the forth attractor, both generators are desynchronized; its basin is yellow. Fig. 3 shows the visualization of the basins of attraction under several values of $Q'_n$.

### 4. Basins of attraction

This section analyzes the basins of attraction of the stable equilibrium points under several fixed values of $Q'_n$. Based on the analysis, we consider the effect of heat transfer management on the dynamics of the electric sub-system. A possible open-loop control is then discussed in terms of a transient instability. Following [6–8], the basin of attraction is visualized by taking a two-dimensional slice of $([\delta_1, \delta_2, \omega_1, \omega_2] \in \mathbb{R}^2 \times \mathbb{R})$ in the entire phase space $X := \mathbb{T}^2 \times \mathbb{R}^2$, where $\mathbb{T}$ stands for the torus, and $\mathbb{R}$ for the set of real number. For the slice, initial conditions on a grid of $401 \times 401$ points were numerically integrated. Each point is colored according to the attractor reached from the corresponding initial condition.
equilibrium points become small on the two-dimensional slices as $Q'_0$ increases. In [2], a similar result is obtained for $D_i = 0.21$ and $\lambda_i = \mu_i = 0$.

This analysis suggests a possibility of instability due to a change of $Q'_0$. In Fig. 3, the solid line shows the stable equilibrium points under various $Q'_a$ between 0 and 0.5, and the dot (●) the equilibrium point under $Q'_a = 0$. The basins of attraction directly illustrates the following two ideal operations of the generators along the lines in Fig. 3. However, a step-wise change of $Q'_0$ from 0 to 5.0 desynchronizes the generator #1 because the dot (●) exists outside the domain of attraction of the stable equilibrium point in Fig. 3d.

As a realistic situation, an open-loop control of the heat transfer rate $Q'_0$ can be considered as in between the above two ideal situations. In this paper, based on [13], we consider a ramp-wise change of the set-points of CHP plants from $Q'_0 = 0$ to 0.5. The duration $T_d$ of the change of the set-points is an important parameter: $T_d = 0$ corresponds to the step-wise change, and $T_d \to \infty$ the quasi-static change. In an engineering viewpoint, the range of $T_d$ where the instability does not occur is of significant importance. Fig. 4 shows the system’s responses for $T_d = 0, 0.5$, and 1.0 s. The red line shows the case of $T_d = 0$ (step-wise change), and the generator #1 is desynchronized as mentioned above. In the case of $T_d = 1.0$ s (blue line), the variables $\delta_i$ and $\omega_i$ converged to the values of the equilibrium point. In the case of $T_d = 0.5$ s, it is observed that the generator #1 is desynchronized. The analysis of the relationship between $T_d$ and the basins of attraction is future work and discussed at the end of this paper.

5. Conclusions and discussion

In this paper, we analyzed the basins of attraction of equilibrium points representing steady operating conditions of synchronous generators in a two-site electricity and heat supply system. The slices of the basins were visualized under various fixed values of heat transfer rate $Q'_0$.

The analysis indicated that the heat transfer management affected the responses of the electric sub-system, and the basin of attraction of the stable equilibrium point became small depending on $Q'_a$. Furthermore, a possibility of instability was discussed for step-wise and ramp-wise changes of the set-points of the CHP plants. It was observed that the instability possibly occurred under a ramp-wise change with a small values of the duration $T_d$.

Finally, for the future work, we discuss the possibility of analyzing $T_d$ via the basins of attraction. After the time $t = T_d$, from the uniqueness of the solution of (6), the resultant behavior is determined by the basins in Fig. 3 if the state trajectory passes the slice determined by $\omega_1 = \omega_2 = 0$. However, in general, this is not the case because the two dimensional slice is not transversal in the full four-dimensional phase space. Nevertheless, it is observed in the case of $T_d = 0.5$ s (green line) in Fig. 4 that the state passes a slice of $(\delta_1, \delta_2, \omega_1, \omega_2) \in X$ for $\omega_i = \epsilon, \omega_2 = 0$ for a small $\epsilon$. Thus, if the basin structure does not vary drastically depending on the values of $\omega_0$, the visualization of the basins in Fig. 3 may be used for the analysis.

References

Existence and stability of odd and even parity discrete breathers in 
Fermi-Pasta-Ulam lattices

Kazuyuki Y oshimura†

†Department of Information and Electronics, Graduate school of Engineering, Tottori University
4-101 Koyama-Minami, Tottori 680-8552, Japan
Email: kazuyuki@eecs.tottori-u.ac.jp

Abstract—Discrete breathers are spatially localized periodic solutions in nonlinear lattices. We have proved the existence of discrete breathers having odd and even parity symmetries, i.e., Sievers-Takeno and Page modes, in one-dimensional Fermi-Pasta-Ulam type lattices for a class of nonhomogeneous potentials. Moreover, we have proved that the Sievers-Takeno mode is spectrally unstable while the Page mode is spectrally stable.

1. Introduction

Spatially localized excitation in nonlinear space-discrete dynamical systems has attracted great interest since the ground-breaking work by Takeno et al. [1, 2]. The localized mode is called discrete breather (DB) or intrinsic localized mode. Considerable progress has been achieved in understanding the nature of DB so far (e.g., [3, 4] and references therein).

The DBs are time-periodic and spatially localized solutions of the equations of motion. From the mathematical point of view, fundamental issues are their existence and stability. The anti-continuous limit is a useful concept for proving the existence of DBs. Existence proofs based on this concept have been given for various lattice models [5, 6]. The stability of DBs also has been studied near the limit [7, 8, 9, 10].

The FPU lattice is one of the fundamental lattice models in physics, to which the anti-continuous limit approach is not applicable. Two types of fundamental DB solutions that have different spatial symmetries, i.e., odd and even parity DB solutions, are known for this model. The odd and even parity DBs are called Sievers-Takeno (ST) mode [1, 2] and Page (P) mode [11], respectively. Normalized spatial profiles of the ST and P modes in a one-dimensional FPU lattice are approximately given by \((\ldots, 0, 1/2, 1, -1/2, 0, \ldots)\) and \((\ldots, 0, -1, 1, 0, \ldots)\) in the regime of strong localization, respectively, provided that interaction potentials of the lattice are of hard type. These two modes were originally found by approximate analytical calculations and then numerically confirmed.

For the FPU model, the first existence proof of DB solutions with odd and even parity symmetries was given in the particular case of homogeneous potential [12]. For more general nonhomogeneous potentials, an existence proof has been given by using a center manifold reduction technique [13]: the existence of DB solutions with odd and even parity symmetries has been proved in a regime of weak localization, where the DBs have small amplitudes and frequencies close to the phonon band edge. In other regimes, no existence proof has been given. As for the stability of DBs, it has been clarified only numerically so far for the FPU lattices [14] and there has been no rigorous result.

In this study, we consider one-dimensional nonhomogeneous potential FPU lattices with periodic boundary conditions, and prove existence of the odd and even parity DBs, i.e., the ST and P modes, in the regime of strong localization. To this end, we develop a new approach which is based on the use of an associated homogeneous potential FPU lattice and Banach’s fixed point theorem. Moreover, we prove that the odd and even parity DB solutions are spectrally unstable and stable, respectively.

2. Lattice model

We consider the one-dimensional FPU lattices described by the Hamiltonian

\[
H = \sum_{i=0}^{N-1} \frac{1}{2} p_i^2 + \sum_{i=0}^{N} V(q_{i+1} - q_i),
\]

where \(q_i \in \mathbb{R}, p_i \in \mathbb{R}, V\) is a potential function, and the periodic boundary conditions \(q_{N+1} = q_0\) and \(p_{N+1} = p_N\) are assumed. Let \(N_0 = 2N + 1\), which represents the number of degrees of freedom. Hamiltonian (1) describes one-dimensional chains of unit-mass particles with nearest neighbour interactions by \(V\). The position and momentum of the \(i\)th particle are represented by \(q_i\) and \(p_i\), respectively.

Let \(X \in \mathbb{R}, \mu \in \mathbb{R}^l\) be a set of parameters, and \(O \subset \mathbb{R}^l\) be a neighbourhood of \(\mu = 0\). We assume the interaction potential \(V\) to be defined by

\[
V(X) = W(X, \mu) + \frac{1}{k} X^4,
\]

where:

(P1) \(k \geq 4\) is an even integer;
(P2) \(W(X, \mu) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}\) is a \(C^2\) function of \(X\) and \(\mu\);
(P3) \(W(X, 0) = 0\) for all \(X \in \mathbb{R}\).
A typical nonhomogeneous potential often used in the literature is polynomial potential. Equation (2) incorporates the polynomial potential \( W(X, \mu) = \sum_{i=1}^{k} \mu_i X^i \), where \( \mu = (\mu_1, \ldots, \mu_k) \), as an example.

Hamiltonian (1) defines the equations of motion in the phase space \( \mathbb{R}^{2N} \) which is endowed with the symplectic 2-form \( \omega = \sum_{i=N-1}^{N} dq_i \wedge dp_i \) as follows:

\[
q_i = p_i, \quad \dot{p}_i = V'(q_{i+1} - q_i) - V'(q_i - q_{i-1}),
\]

where \( i = -N, \ldots, N \). Let \( \Gamma(t) = (q(t), p(t)) \in \mathbb{R}^{2N} \) be a \( T \)-periodic solution of Eq. (3), where \( q = (q_{-N}, \ldots, q_N) \) and \( p = (p_{-N}, \ldots, p_N) \) are the position and momentum vectors. Let \( \xi \) be the variation in \( q \), and we use the notation \( \xi = (\xi_{-N}, \ldots, \xi_N) \). Linearizing Eq. (3) along \( \Gamma(t) \), we obtain the variational equations in the second-order differential equation form as follows:

\[
\ddot{\xi} + A(t) \xi = 0, \tag{4}
\]

where \( A(t) \) is the Hessian matrix of the potential function evaluated on \( \Gamma(t) \), i.e., its components are given by \( [A(t)]_{ij} = \partial^2 U(q(t)) \partial q_i \partial q_j \), where \( U = \sum_{i=-N}^{N} V(q_{i+1} - q_i) \).

Let \( \{\xi_1, \ldots, \xi_{2N}\} \) be a system of fundamental solutions of Eq. (4). According to the Floquet theory, the fundamental solutions of Eq. (4) at \( t \) and \( t + T \) are related via a \( 2N_0 \times 2N_0 \) monodromy matrix \( M \) as

\[
(\xi_1(t + T), \ldots, \xi_{2N}(t + T)) = (\xi_1(t), \ldots, \xi_{2N}(t)) M. \tag{5}
\]

Eigenvalues of \( M \) are called the characteristic multipliers and they are independent of the choice of fundamental solutions. Let \( p_i, i = 1, \ldots, 2N_0 \) be the characteristic multipliers of \( \Gamma(t) \). The spectral stability of \( \Gamma(t) \) is defined as follows.

**Definition 1.** Periodic solution \( \Gamma(t) \) is said to be spectrally unstable if there exists \( p_i \) such that \( |p_i| > 1 \), otherwise it is said to be spectrally stable.

### 3. Symmetry of solution

We precisely describe the odd and even parity symmetries in this section. Let \( S_O \) and \( S_E \) be the linear mappings \( S_O, S_E : \mathbb{R}^N \to \mathbb{R}^N \) defined by

\[
S_O : (S_O \cdot x)_i = -x_i, \quad i = -N, \ldots, N, \\
S_E : (S_E \cdot x)_i = -x_{i+1}, \quad i = -N, \ldots, N,
\]

where \( x = (x_{-N}, \ldots, x_N) \in \mathbb{R}^N \) represents a point in the space \( \mathbb{R}^N \) and \( x_{-(N+1)} = x_N \) due to the periodic boundary conditions. These \( S_O \) and \( S_E \) are linear involutions, i.e., \( S_O \circ S_O = S_E \circ S_E = id \).

Let \( \Gamma(t) = (q(t), p(t)) \in \mathbb{R}^{2N_0} \) denote a periodic solution of Eq. (3) with a period \( T \). The solution \( \Gamma(t) \) is said to have odd symmetry if it satisfies the relations

\[
S_O q(t + T/2) = q(t), \quad S_O p(t + T/2) = p(t), \quad \forall t \in \mathbb{R}. \tag{6}
\]

When the interaction potential is an even function, i.e., \( V(X) = V(-X) \), an additional symmetry \( \Gamma(t + T/2) = -\Gamma(t) \) holds. Then equation (6) reduces to

\[
-S_O q(t) = q(t), \quad -S_O p(t) = p(t), \quad \forall t \in \mathbb{R}. \tag{7}
\]

On the other hand, \( \Gamma(t) \) is said to have even symmetry if it satisfies the relations

\[
S_E q(t) = q(t), \quad S_E p(t) = p(t), \quad \forall t \in \mathbb{R}. \tag{8}
\]

Equations (6) and (8) correspond to the solution profiles centered at \( i = 0 \) site and that centered between \( i = -1 \) and \( 0 \) sites, respectively.

### 4. Notations

We introduce some notations to state the main theorems. Consider the scalar differential equation

\[
\phi + \phi^{k-1} = 0. \tag{9}
\]

Equation (9) has the energy integral \( \phi^2/2 + \phi^k/h = h \), where \( h \) is an integration constant. Its solution is non-constant and periodic for any given \( h > 0 \). Let \( \phi(t) \) be the solution of Eq. (9) with initial conditions \( \phi(0) = (kh)^{1/k} > 0 \) and \( \phi(0) = 0 \). The period \( T \) of \( \phi(t) \) depends on \( h \), and it is obtained from the energy integral as follows:

\[
T = 2 \sqrt{2} h^{-1/(2-1/k)} \int_0^{T/2} \frac{1}{\sqrt{1 - x^2/k}} dx. \tag{10}
\]

This indicates \( T \propto h^{-1/(2-1/k)} \) and that \( T \) monotonically decreases from \( T = +\infty \) to 0 as \( h \) varies from \( h = 0 \) to \( +\infty \), since the integral in Eq. (10) is independent of \( h \). Thus, for any given \( T > 0 \), there exists a non-constant periodic solution \( \phi(t) \) with the period \( T \), which corresponds to a value of \( h \) uniquely determined from Eq. (10). We denote this \( T \)-periodic solution of Eq. (9) with \( \phi(t; T) \).

Let \( \Pi_O \) and \( \Pi_E \) be the subspaces of \( \mathbb{R}^{N_0} \) defined by

\[
\Pi_O = \left\{ x \in \mathbb{R}^N \mid S_O \cdot x = x \right\}, \tag{11}
\]

\[
\Pi_E = \left\{ x \in \mathbb{R}^N \mid S_E \cdot x = x \right\}, \tag{12}
\]

where \( x = (x_{-N}, \ldots, x_N) \). These \( \Pi_O \) and \( \Pi_E \) are subspaces in the configuration space \( \mathbb{R}^N \) that satisfy the odd and even symmetries, respectively (cf. Eqs. (7) and (8)).

Let \( m \in \mathbb{N}, c > 0, 0 < r < 1 \) be parameters. We define a closed subset \( B_{m,r,c} \subset \mathbb{R}^{N_0} \) as follows:

\[
B_{m,r,c} = \left\{ x \in \mathbb{R}^N \mid |x| \leq c \text{ for } 0 \leq i \leq m, \right. \\
\left. |x_i| \leq cr^{(k-1)-r} \text{ for } m + 1 < i \leq N \right\}. \tag{13}
\]

This subset \( B_{m,r,c} \) is specified by the three parameters \((m, r, c)\). Equation (13) shows that the interval of \( x_i \) rapidly decreases with increasing \( i \) in \( B_{m,r,c} \). 

- 680 -
Consider the phase space $\mathbb{R}^{2N}$ of Hamiltonian system (1). Let $\Pi$ be the subspace of $\mathbb{R}^{2N}$ defined by

$$
\Pi = \left\{ (q, p) \in \mathbb{R}^{2N} : \sum_{i=-N}^{N} g_i = \sum_{i=-N}^{N} p_i = 0 \right\}.
$$

This is the subspace in which both the mass center and the total momentum are zero. Since $d(\sum_{i=-N}^{N} p_i)/dt = 0$ follows from Eq. (3) and the periodic boundary conditions, $\Pi$ is an invariant subspace of Hamiltonian system (1).

5. Main results

Our main theorems for the existence and spectral stability are stated as follows. Theorems 1 and 2 are for the odd and even parity DB solutions, i.e., the ST and P modes, respectively.

**Theorem 1.** Suppose potential function (2) and (P1)-(P3). Let $\{a_i\}_{i=-N}^{N}$ and $(m, c, r)$ be constants given in Table 1 for $k$. Then, for any $N \geq 4$ and any $T > 0$, there exists a unique $x \in B_{m,c,r} \cap \Pi_{E}$ such that $\Gamma_{ST}(t; T) = (u(x; t), u(x; T)) \in \mathbb{R}^{2N}$ is a $T$-periodic solution of FPU lattice (1) with $\mu = 0$, where $u = a + x$ and $a = (a_{-N}, \ldots, a_N)$. Moreover, there exist a neighbourhood $U \subset \mathbb{R}$ of $\mu = 0$ and a unique family $\Gamma_{ST}(t; T, \mu)$ of $T$-periodic solutions of FPU lattice (1) for $\mu \in U$ such that $\Gamma_{ST}(T; T, \mu) = \Gamma_{ST}(0; T, \mu)$, and it satisfies odd symmetry (6). The periodic solution $\Gamma_{ST}(t; T, \mu)$ is spectrally unstable with one unstable characteristic multiplier.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$a_7$</th>
<th>$a_8$</th>
<th>$a_9$</th>
<th>$a_{10}$</th>
<th>$a_{11}$</th>
<th>$a_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 4$</td>
<td>0.3762</td>
<td>-0.1968</td>
<td>8.67 × 10⁻³</td>
<td>0</td>
<td>(m, c, r) = (3.9 × 10⁻³, 3 × 10⁻³)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 6$</td>
<td>0.5057</td>
<td>-0.2539</td>
<td>1.1 × 10⁻³</td>
<td>0</td>
<td>(m, c, r) = (3.8 × 10⁻⁴, 7 × 10⁻⁴)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k \geq 8$</td>
<td>2 × 3^(-(k-1)/(k-2))</td>
<td>-3^(-(k-1)/(k-2))</td>
<td>0</td>
<td>(m, c, r) = (2.02 × 3^(-(k-1)/(k-2)), 5 × 10⁻³)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1

**Theorem 2.** Suppose potential function (2) and (P1)-(P3). Let $\{a_i\}_{i=-N}^{N}$ and $(m, c, r)$ be constants given in Table 2 for $k$. Then, for any $N \geq 4$ and any $T > 0$, there exists a unique $x \in B_{m,c,r} \cap \Pi_{E}$ such that $\Gamma_{ST}(t; T) = (u(x; t), u(x; T)) \in \mathbb{R}^{2N}$ is a $T$-periodic solution of FPU lattice (1) with $\mu = 0$, where $u = a + x$ and $a = (a_{-N}, \ldots, a_N)$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$a_0$</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
<th>$a_5$</th>
<th>$a_6$</th>
<th>$a_7$</th>
<th>$a_8$</th>
<th>$a_9$</th>
<th>$a_{10}$</th>
<th>$a_{11}$</th>
<th>$a_{12}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$k = 4$</td>
<td>0.323</td>
<td>-0.0535</td>
<td>0</td>
<td>(m, c, r) = (2.3 × 10⁻⁴, 6 × 10⁻³)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 6$</td>
<td>0.4166</td>
<td>-0.015</td>
<td>0</td>
<td>(m, c, r) = (2.9 × 10⁻⁴, 7 × 10⁻⁴)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 8$</td>
<td>0.4485</td>
<td>-3.65 × 10⁻³</td>
<td>0</td>
<td>(m, c, r) = (2.2 × 10⁻⁴, 2 × 10⁻⁴)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k = 10$</td>
<td>0.45839</td>
<td>-9.1 × 10⁻⁴</td>
<td>0</td>
<td>(m, c, r) = (2.9 × 10⁻⁴, 8 × 10⁻⁴)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k \geq 12$</td>
<td>(1 + 2^(-1))^1/(k-2)</td>
<td>0</td>
<td>(m, c, r) = (2.02(1 + 2^(-1))^1/(k-2), 2 × 10⁻³)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2

**Remark 1.** If $\Gamma(t)$ is a periodic solution of autonomous Hamiltonian system (1), then so is a phase-shifted solution $\Gamma(t + r)$ for $r \in \mathbb{R}$ and it has the same orbit as $\Gamma(t)$ in the phase space. Two periodic solutions are identified if they differ only by a phase-shift.

**Remark 2.** We assumed the $\mu$-dependent potential $W$ of nearest-neighbor interaction type in Eqs. (1) and (2), focusing on the FPU lattices. This assumption is not essential. The statements of Theorems 1 and 2 hold for a more general Hamiltonian $H = \sum_{i=-N}^{N} p_i^2/2 + \sum_{i=-N}^{N} (q_{i+1} - q_i)^2/k + W(q, \mu)$, provided that $W$ is a $C^2$ function of $q$ and $\mu$ such that $W(q, 0) = 0$ and $W(q + \epsilon, \mu) = W(q, \mu)$ for all $\epsilon \in \mathbb{R}$, where $\epsilon = (1, \ldots, 1) \in \mathbb{R}^N$.

**Remark 3.** Theorems 1 and 2 imply the existence of spatially periodic array of odd or even parity DB solutions in the infinite FPU lattices.

In Theorem 1, the approximation vector $a$ has non-zero components on only a small number of sites with the indices satisfying $|i| < m$, and it represents a strongly localized profile. The theorem states that the profile vector $u$ of
$\Gamma_\text{ST}(t; T)$ is close to $a$ and satisfies

$$|u_i - a_i| \leq \begin{cases} \frac{c}{r^{k-1-m}} & \text{if } |i| \leq m, \\ c r (k-1) |i|^{-m} & \text{otherwise.} \end{cases}$$

In addition, the latter inequality indicates rapid decrease of $|u_i|$ with increasing $|i|$, since it is equivalent to $|u_i| \leq c r^{k-1-m}$ due to $a_i = 0$, $|i| \geq m$. Thus $\Gamma_\text{ST}(t; T)$ is a strongly localized solution. The solution $\Gamma_\text{ST}(t; T; \mu)$ is also strongly localized for small $\mu$ because of its continuity with respect to $\mu$. Similarly, Theorem 2 shows that both $\Gamma_\text{ST}(t; T)$ and $\Gamma_\text{FPU}(t; T; \mu)$ have strongly localized profiles. The approximations $\{a_i\}_{i=-N}^{N}$ play a crucial role in the theorems.

6. Strategy of proofs of the theorems

We outline our strategy for proving Theorems 1 and 2. First, we consider the homogeneous potential FPU lattice which is described by Hamiltonian (1) with the potential $V(X) = X^4/k$, i.e., $\mu = 0$ in Eq. (2). In this particular lattice, it is possible to find a DB solution in the form $u = ub(t)$, where $u \in \mathbb{R}^N$ is a constant vector describing the spatial profile of the solution. The problem of finding a DB solution is reduced to a set of algebraic equations for $u$, and we solve it by using Banach’s fixed point theorem in a neighborhood of the approximation $\{a_i\}_{i=-N}^{N}$. This fixed point approach enables one to obtain a precise quantitative estimation of $u$ that has odd or even parity symmetry. Using this estimation of $u$, we evaluate the characteristic multipliers, i.e., the spectral stability, of the DB solution.

Next, we consider the nonhomogeneous potential FPU lattice, i.e., $\mu \neq 0$ in Eq. (2). The DB solution in the homogeneous potential lattice is continued to a nonhomogeneous potential one for small $\mu \neq 0$ by using the implicit function theorem, based on the characteristic multipliers for $\mu = 0$. We show that this continuation retains odd or even parity symmetry of the DB solution. Finally, we evaluate variations of the characteristic multipliers under the perturbation in $\mu$ to determine the spectral stability of the DB solution for $\mu \neq 0$.

References


Generalised Weierstrass elliptic functions and nonlinear wave equations

J. Chris Eilbeck†

†Maxwell Institute and Department of Mathematics, Heriot-Watt University, Edinburgh, Scotland, EH14 4AS, UK
Email: J.C.Eilbeck@hw.ac.uk

Abstract—The well-known Weierstrass elliptic functions are constructed from an algebraic curve of genus \( g = 1 \), and can be used to solve a number of nonlinear ordinary differential equations, such as the travelling wave
problem for the KdV equation. As well as the soliton solution, such methods give periodic solutions of the ODEs. If the curve is generalised to a higher genus, the corresponding generalised Weierstrass functions give multiple periodic solutions of many well-known PDEs, such as the
KdV equation \((g = 2)\), the Boussinesq equation \((g = 3)\), and the Kadomtsev-Petviashvili (KP) equation \((g = 6)\). We review, very briefly, some of the results in this area.

1. Introduction

The well-known Weierstrass elliptic functions were first studied in the middle of the 19th century by Abel, Hermite and Weierstrass. The functions were related to an elliptic curve, given in standard Weierstrass form as
\[
y^2 = 4x^3 - g_2 x - g_3,
\]
where \( g_2 \) and \( g_3 \) are parameters. We refer to this as the \((2,3)\) curve. If \( x \) and \( y \) are considered to be complex variables, the resulting surface has the shape of a torus, which has genus \( g = 1 \). Associated with the curve \((1)\) is the Weierstrass \( \wp \) function, where the variable \( u \) satisfies \( u = \int dx/y \), where \( dx/y \) is called a differential of the first kind. The doubly periodic function \( \wp(u) \) satisfies the ODEs
\[
(\wp')^2 = 4\wp^3 - g_2 \wp - g_3, \quad (2)
\]
\[
\wp'' = 6\wp^2 - 1/2g_2. \quad (3)
\]
The second ODE follows easily from the first after differentiation and cancellation. Additionally, we have the Weierstrass \( \sigma(u) \) function, which is related to \( \wp \) by
\[
\wp(u) = -\frac{d^2}{du^2} \ln \sigma(u). \quad (4)
\]
A well-known application of \((2)\) is to the travelling wave solution of the Korteweg-de Vries (KdV) equation
\[
u_t + 6u \nu_x + \nu_{xxx} = 0, \quad (5)
\]
where the subscripts denote partial differentiation. Putting
\[u(x,t) = f(x-ct) = f(\zeta),\]
we find
\[
-\frac{df}{d\zeta} + 6f \frac{df}{d\zeta} + \frac{d^3f}{d\zeta^3} = 0. \quad (6)
\]
Integrating with respect to \( \zeta \) we get, on setting the constant of integration equal to zero
\[
\frac{d^2f}{d\zeta^2} + 3f^2 - cf = 0, \quad (7)
\]
which is just \((3)\) with \( f = -2\wp(u) + c/6, g_2 = c^2/12 \).

We discuss below what happens when we generalise the curve \((1)\) to higher genus curves, with powers of \( x \) and \( y \) greater than \( 2 \) or \( 3 \) respectively. We can then find generalisations of the ODEs \((2)\) and \((3)\) which are now PDEs in functions of \( g \) variables.

2. Genus 2 case

In genus 2 the relevant curve is hyperelliptic (leading term in \( y \) is \( y^3 \)) \((2,5)\) curve
\[
C : \quad y^2 = x^5 + \mu_2 x^4 + \mu_3 x^3 + \ldots + \mu_0. \quad (8)
\]
This case was considered in detail by Baker (1907). \( \sigma \) and \( \wp \) are now functions of \( g = 2 \) variables, i.e.
\[
\sigma = \sigma(u_1, u_2) = \sigma(\mu)
\]
There are now two differentials of the first kind, \( dx/y \) and \( x dx/y \), and we define \( u_1, u_2 \) by
\[
\begin{align*}
  u_1 &= \int_{x_1, y_1}^{x_2, y_2} \frac{dx}{y}, \\
  u_2 &= \int_{x_1, y_1}^{x_2, y_2} \frac{x dx}{y},
\end{align*}
\]
for two variable points \((x_1, y_1)\) on \( C \).

2.1. Genus 2 PDEs

Now that our function sigma depends on several variables, we need a new notation corresponding to the 2nd logarithmic derivative in \((4)\). The generalized \( \wp \) functions (note more than 1 type!) are defined from the \( \sigma \) function as
\[
\wp_{ij}(u_1, u_2) \equiv -\frac{\partial^2}{\partial u_i \partial u_j} \ln \sigma(u_1, u_2), \quad i, j = 1, 2. \quad (11)
\]
There is a nice if somewhat complicated expansion formula due to Klein to get the PDE’s, once the curve and the
differentials have been fixed. In the genus 2 case this gives
\[ \wp_{2222} - 6\wp_{22}^2 = 2\mu_4 + 4\mu_2\wp_{22} + 4\wp_{12}, \]
\[ \wp_{1122} - 2\wp_{22}\wp_{11} = 4\wp_2\wp_{12} - 2\wp_{11}, \]
\[ \wp_{1111} - 6\wp_{11}^2 = 4\wp_2\wp_{12} - 2\wp_8\wp_{22} - 4\mu_{10}, \]
\[ -12\mu_{10}\wp_{22} - 2\mu_{10}\wp_2 - 2\mu_8\wp_{12}. \]

These are the generalization of \( \wp'' - 6\wp^2 = -1/2g_2 \) in genus one (\( \wp_{1111} - 6\wp_{11}^2 = -1/2g_2 \) in our new PDE notation).

If we take the first of these equations, differentiate once w.r.t. \( u_2 \), then put \( u_1 = t, u_2 = x, \wp_{22} = -U(x,t), \mu_2 = 0 \), we find
\[ U_t + 12U_{tx} + U_{xxx} = 0, \]
which is the KdV equation. Our function \( \wp_{22} \) is hence a fully two-dimensional multi-periodic solution of the KdV equation. This result was derived by Baker in 1907 [1], some 60 years before the discovery of the “modern” multi-soliton solutions of KdV.

The \( \sigma \) function in the genus 2 case can be expanded in a power series in \( u_1, u_2 \)
\[ \sigma(u_1, u_2) = u_1 - \frac{1}{5} u_3^3 - \frac{1}{60} \mu_2 u_2^5 \]
\[- \frac{1}{1260} (8u_2^2 + \mu_2) u_2^7 + \frac{1}{12} \mu_4 u_1 u_2^4 + \ldots \]
The function \( \sigma \) satisfies four linear heat-type equations in the \( u_i \) and the \( \mu_j \), the first of which is
\[ 4\mu_4 \frac{\partial \sigma}{\partial u_4} + 6\mu_6 \frac{\partial \sigma}{\partial u_6} + 8\mu_8 \frac{\partial \sigma}{\partial u_8} + 10\mu_{10} \frac{\partial \sigma}{\partial u_{10}} = -3u_1 \frac{\partial \sigma}{\partial u_1} + u_2 \frac{\partial \sigma}{\partial u_2} - \sigma - 3 \]
These heat-type equations can be used to form recurrence relations satisfied by the coefficients of the sigma expansion.

Many results for the hyperelliptic cases \( y^2 = x^3 + \ldots \), including many general theorems, have been derived in [2].

3. Genus 3

Here we consider only the so-called trigonal genus 3 curve, which is non-hyperelliptic [3]
\[ C : y^3 = x^4 + \mu_3 x^3 + \ldots + \mu_0, \]
Now all functions are functions of \( u = (u_1, u_2, u_3) \), defined in an analogous way to (9). The resulting \( \wp \) functions coming from this (3,4) curve satisfy a number of PDEs, the first of which is
\[ \wp_{333} - 6\wp_{33}^2 = -3\wp_{23}, \]
which can be shown to transform to the Boussinesq equation
\[ U_t - U_{xx} - \frac{\partial}{\partial x} \left( 3u^2 + U_{xx} \right) = 0 \]
The corresponding \( \sigma \) function can again be shown to solve a set of heat-type equations, which can also be used to form a recurrence relations for the coefficients of the corresponding sigma expansion.

Similar results have been derived for higher genus trigonal curves \( y^3 = x^3 + \ldots, y^3 = x^3 + \ldots \), etc.

4. Genus 6

In genus 6 a new possibility occurs, the so-called tetragonal (4,5) curve [4]
\[ y^4 = x^5 + \ldots. \]
Many of the corresponding equations have been worked out in this case, although things get quite complicated. Now the \( \wp \) and \( \sigma \) functions are functions of six variables \( u_i, i = 1, \ldots, 6 \). The \( \wp \) functions satisfy a number of PDEs, the first of which is
\[ \wp_{6666} - 6\wp_{66}^2 = 3\wp_{55} + 4\wp_{46}. \]  
(12)
Differentiating (12) twice with respect to \( u_6 \) we get
\[ \wp_{6666} = 12 \frac{\partial}{\partial u_6} \wp_{66} \wp_{66} - 3\wp_{55} + 4\wp_{46}. \]
Making the substitutions \( u_6 = x, u_5 = y, u_4 = t, \) and \( U(x,y,t) = \wp_{66} \), we get eventually the following form of the KP-equation:
\[ (U_{xxx} - 12U_{xx}U_t - 4U_t)_x + 3U_{yy} = 0. \]

Acknowledgements

This is work done in collaboration with a number of co-authors, including Yoshiiro Ōnishi of Meijo University, Nagoya, and John Gibbons of Imperial College, London.

References

Mobility of Discrete Breather in Truncated Pairwise Interaction Symmetric Lattices

Yusuke Doi$^\dag$ and Kazuyuki Yoshimura$^\ddag$

$^\dag$Department of Adaptive Machine Systems, Graduate School of Engineering, Osaka University
2−1 Yamadaoka, Suita, Osaka 565-0871, Japan
$^\ddag$Department of Information and Electronics, Graduate School of Engineering, Tottori University
4-101 Koyama-Minami, Tottori 680-8552, Japan
Email: doi@ams.eng.osaka-u.ac.jp, kazuyuki@eecs.tottori-u.ac.jp

Abstract—Traveling discrete breathers (DBs) in an approximated version of the pairwise interaction symmetric lattice (PISL) are investigated. The approximated PISL is constructed by truncating the pairwise interactions up to a small number of nearest neighbor particles. Even in the truncated PISL, traveling DBs have good mobility. An effective periodic barrier for the traveling DB becomes smaller than that of the original Fermi-Pasta-Ulam $\beta$ lattice.

1. Introduction

Mobility of discrete breather (DB) is one of the most important problems in nonlinear lattice dynamics since traveling DBs can transport energy in discrete systems such as crystals, metamaterials and MEMS. The traveling DBs are highly affected by the discreteness of the system for example, the discreteness causes reduction and variation of velocity of DB. We have proposed a new lattice model called a pairwise interaction symmetric lattice (PISL) that counteracts the effects of discreteness. The PISL has all-to-all connection between particles. However, it is difficult to realize this all-to-all connection in practical physical systems. In this study, we propose an approximated PISL that may reduce this difficulty.

2. Model

We consider an nonlinear lattice model with $N$ particles. Hamiltonian of the system is

$$H = \sum_{n=1}^{N} \frac{p_n^2}{2} + \sum_{n=1}^{N} \frac{1}{2} (q_{n+1} - q_n)^2 + \sum_{n=1}^{N} \frac{\beta}{4} (q_{n+1} - q_n)^4 + \sum_{n=1}^{N} \sum_{d=2}^{M} \frac{\alpha_d}{4} (q_{n+d} - q_n)^4,$$

where $p_n$ and $q_n$ are the momentum and displacement of $n$th particle, respectively, $\beta$ is the coefficient for nonlinear interaction between nearest neighbor particles, $\alpha_d$ is the coefficient for nonlinear interaction between $d$th nearest neighbor particles which is calculated according to the procedure for constructing the PISL [2]. In Eq.(1), the interactions up to $M$th nearest particles are considered. The system reduces to the Fermi-Pasta-Ulam (FPU) $\beta$ lattice when $\alpha_d = 0$ for all $d \geq 2$. The system becomes the original PISL when $M = N/2$.

3. Numerical Results

Figure 1 shows the trajectory of the traveling DBs in $(X,\dot{X})$ space.

![Figure 1: Trajectory of traveling DBs in $(X,\dot{X})$ space.](image)

Figure 1 shows the trajectory of the traveling DBs in $(X,\dot{X})$ space with $M = 1, 2, 5, 10$ and 20. The parameters $X$ and $\dot{X}$ are the short time averaging position and velocity of the DB, respectively.

In the case of the original FPU-$\beta$ lattice, a large variation in the velocity of DB is observed. By introducing the long range interactions, the variation of velocity decreases. The variation of velocity indicates that the existence of an effective potential barrier for the traveling DB. Figure 1 shows that the effective potential barrier can be reduced by considering the truncated PISL. This result indicates that the truncated PISL supports the smooth mobility of traveling DB and is useful for constructing the practical physical systems that supports the traveling DBs.

References


Experimental Excitations of Intrinsic Localized Modes in an Air-Levitation-Type Coupled Oscillator Array

Yosuke Watanabe† and Mai Nishimoto†

†Department of Mechanical Science, Graduate School of Engineering Science, Osaka University
Toyonaka, Osaka 560-8531, Japan
Email: watanabe@me.es.osaka-u.ac.jp (Watanabe)

Abstract—An air-levitation-type coupled oscillator chains has been constructed with improving nonlinear springs. The relation of restoring force of the spring to deflection is approximately cubic and symmetrical with respect to the equilibrium in the apparatus with appropriate tensions. It has been observed that, driving one end of the chains sinusoidally with a frequency above the cutoff, localized oscillations can be excited intermittently at the driven end and they are propagated along the chains at a constant speed.

1. Introduction

It is now known that the intrinsic localized modes (ILMs) or the discrete breathers (DBs) are generic in spatially periodic, discrete and nonlinear systems. It is also known that the mobile type of ILMs can be excited both in a spatially infinite system and a semi-infinite system driven at one end sinusoidally at a frequency in a linear stopping band above the passing band (for example [1]). We have shown numerically that ILMs can be excited and propagated in the finite mass-spring chains with the piecewise linear springs driven sinusoidally at one end with the other fixed [2, 3]. The relation between the driving amplitude and frequency to excite ILMs is similar to the one for the ILMs in the case of the FPU-β model unless the amplitude is too large. The results of numerical studies should be used to implement the experiments and we have actually succeeded in experimental demonstrations of ILMs. (The results of the experiments will be published in a forthcoming paper.) We have constructed, improved several mechanical apparatuses of mass-spring chains and take on the experimental studies on ILMs. In this paper, we report manufacturing air-levitation-type chains and the experimental results.

2. Experiments

We have realized the piecewise linear (nonlinear) springs by using and processing barrel-shaped drawn coil springs. In forming a coil spring, previously decided initial tension is provided to spring to change its hardness (i.e. spring constant) to extension. Moreover between two separate positions in the coil, light and non-stretchable strings are stretched to partially suppress elongation of the spring to extension. By these two processing and embedding in the chains with an appropriate tension, each spring could behave nonlinearly to displacements and symmetrically with respect to its quiescent state.

The experimental apparatus we made consists of twenty identical oscillators, nonlinear springs, straight air track (hollow square pipe with many fine air holes) with a blower and driver for forcing the chains at an end. The chains of oscillators connected with neighboring ones through the springs are levitated above the track by pressurized-air shots from holes by the blower, therefore, the oscillators can be moved linearly with less friction.

3. Results

To excite modes of oscillations in the chains, one end of the chains is driven sinusoidally in the direction of the chains at a frequency, whereas the other is fixed. It has been observed that, driving with a frequency above the cutoff, localized oscillations can be excited intermittently at the driven end and they are propagated along the chains at a constant speed. The FFT shows that the peak of the spectrum of the oscillations is located in the region where linear wave propagations are prohibited. Therefore, the localized oscillations can be regarded as the mobile type of ILMs. It has been also shown that, the driving with higher frequency, the interval between excitations of localized oscillations is increased and they are no longer excited when the frequency exceeds a threshold.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP24654124, JP26400394.

References

Existence Condition and Stability of Rotating Intrinsic Localized Modes in FPU-β Chain with Fixed Boundaries

Atsuki Mitani†, Masayuki Kimura‡, Shinji Doi

†‡Department of Electrical Engineering, Kyoto University, Katsura, Nishikyo-ku, Kyoto 615-8510, Japan
Email: †mitani.atsuki@rotary.kuee.kyoto-u.ac.jp, ‡mkimura@kuee.kyoto-u.ac.jp

Abstract—Intrinsic localized mode (ILM) is a spatially localized and temporary periodic solution in nonlinear lattices. In this report, Rotating ILMs in one-dimensional Fermi-Pasta-Ulam (FPU) chain placed in three-dimensional space are focused on. First we derive an equation of motion and dispersion relations of longitudinal and transverse waves. By using Newton-Raphson method, we investigate existence regions of the rotating ILMs in two-dimensional parameter space consisting of the rotational period and the initial extension of the chain. In addition, stability is evaluated for the rotating ILM in the regions by using characteristic multipliers. As a result, It is shown that the rotating P mode exist in wider region than ST mode. Moreover, we derive the upper bound of the rotational period in the case of P mode.

1. Introduction

It is well known that a spatially localized and temporary periodic solution called intrinsic localized mode (ILM) exist in nonlinear lattice[1], such as graphene, protein, and DNA[2, 3]. It implies that ILM can be utilized to nanotechnology. We have focused on the carbon monoatomic chain in which carbon atoms are linearly-arranged. It has been reported that the carbon monoatomic chain can be fabricated from graphene sheet[4] and it will have very high heat conductivity at the room temperature[5]. Neighboring carbon atoms nonlinearly interact each other with respect to the distance between them[6]. Therefore, vibration of each carbon atom in the chain can be modeled as a nonlinear lattice. The aim of our research is to find ILM in the carbon monoatomic chain and to utilize it to control the heat conduction.

Although the carbon monoatomic chain can be modeled as a nonlinear lattice, there is the significant difference that each carbon atom can move not only the axial direction but also the radial direction. The degree of freedom of the radial direction will cause many differences from the traditional one-dimensional nonlinear lattice that the motions are constrained along the axis of the lattice. Then, we have first focused on the Fermi-Pasta-Ulam-β (FPU-β) chain which is well-known model in which ILM exists[7]. In our previous research, two novel types of ILM was found, namely, the transverse and the rotating ILMs[8]. In this paper, stability and existence parameter regions of the rotating ILMs are investigated.

2. FPU-β chain in three dimensional space

Fermi-Pasta-Ulam chain is the one-dimensional nonlinear lattice that has linear and cubic interaction. Equilibrium state of the chain is shown Fig.1(a). In the figure, l0 is natural length of nonlinear spring connecting neighboring masses and a is displacement. If a is positive, the chain is initially extended, and vice versa. When the system is placed in three-dimensional space, any masses can move not only longitudinally (along the axis) but also transversely (perpendicular to the axis) (see Fig.1(b)). Therefore, the equation of motion of the masses is described as:

\[
\frac{d^2 r_n}{dt^2} = \alpha ((r_{n+1} - r_n) - l_0) \frac{r_{n+1} - r_n}{|r_{n+1} - r_n|} + \beta ((r_{n+1} - r_n) - l_0) \frac{r_{n+1} - r_n}{|r_{n+1} - r_n|} + 
\]

where \( r_n \) is the position vector from equilibrium position to each mass. \( \alpha \) and \( \beta \) are coefficients of the linear and cubic nonlinear interaction, respectively. In this paper, \( r_n \) is treated as a three-dimensional vector in the Cartesian coordinate system, namely, \( r_n = (n(l_0 + a) + x_n, y_n, z_n) \) where \( x_n \) is longitudinal displacement and \( y_n \) or \( z_n \) is transverse displacement of each mass.

Dispersion relation can be obtained by linearizing Eq.(1)

\[
\omega_x = 2 \sqrt{\frac{\alpha + 3\beta a^2}{m}} \sin \frac{k_z}{2}, \quad (2)
\]

\[
\omega_y = 2 \sqrt{\frac{\alpha a + \beta a^3}{m(l_0 + a)}} \sin \frac{k_z}{2}, \quad (3)
\]

where \( \omega_x \) and \( \omega_y \) denote angular frequency of longitudinal and transverse waves, respectively. Since the chain is cylindrically symmetric, \( \omega_z \) is the same as \( \omega_x \). Eqs.(2) and (3) show that there are upper bound for the frequencies \( \omega_x \) and \( \omega_y \). To exist an ILM, the ILM should avoid to resonate with
linear plain waves, namely, the frequency of the ILM $\omega_b$ should be greater than $\omega_x = 2\sqrt{\frac{\alpha^2 + \beta^2}{m}}$ or $\omega_y = 2\sqrt{\frac{\alpha^2 + \beta^2}{m(\rho + \alpha^2)}}$.

In this paper, $m$, $\alpha$, and $l_0$ are set unity, and $\beta$ is fixed at 25. The number of masses is chosen to be eight.

3. Rotating ILM

In the FPU-$\beta$ chain placed in three-dimensional space, three kind of ILM exist, namely, longitudinal, transverse, and rotating ILMs (see Fig.2)[8]. The longitudinal ILM coincides with the traditional ILM in which each mass move only along the axis of the chain. On the other hand, transverse ILM mainly consists of perpendicular oscillation to the axis. For more details of these two types of rotational period of ILMs, see Ref.[8]. The rotating ILM which is focused on in this paper is also a spatially localized and temporary periodic solution of Eq.(1) in which each mass rotates around the axis of the chain. An example of the rotating ILM is shown Fig.3. As shown in the Fig.3, a few masses have large amplitude and the phase difference between $y_n$ and $z_n$ equal $\pi/2$. Thus, each mass rotates around the axis, and the radii of the rotations are localized. In addition, in Fig.3(a), two neighboring masses (4th and 5th) oscillate in antiphase with the same amplitude while the other masses show very small amplitude motion. Therefore, the amplitude distribution correspond to Page mode[7]. In this paper, rotating P mode is abbreviated as R-P mode. Fig.3(b) shows a different ILM from R-P mode in spatial symmetry of the amplitude distribution. The 4th mass has the largest amplitude and the 3rd and 5th masses have the same amplitude which rather smaller than that of the 4th mass but sufficiently large comparing with the other masses. This amplitude distribution correspond to Sievers-Takeno mode[7]. The rotating Sievers-Takeno mode is also abbreviated as R-ST mode. Note that, the length of all the nonlinear springs connecting nearest neighbor masses are time-invariant while the rotating ILM is oscillating. That is, all the masses do not oscillate with changing the length of the nonlinear springs like the longitudinal and transverse ILMs. Therefore, another existence condition should be considered instead of the non-resonant condition Eqs.(2) and (3).
4. Analysis of rotating ILM

In this section, we focus on the rotating ILMs mentioned in the previous section. Stability and existence region of the rotating ILM will be discussed below.

4.1. Dependence on parameter and bifurcation

Figure 4(a) shows R-P modes with respect to the rotation period when the initial extension \(a\) is fixed at 0.1. In the figure, the vertical axis shows the maximum radius of the R-P mode and horizontal axis shows the rotation period. The radius of the R-P mode decreases as the rotation period increases and disappears at about \(T = 9.5\) through a bifurcation. On the other hand, as shown in Fig.4(b), the R-ST mode disappear at about \(T = 4.7\), at which two branches coalesce. The other branch consist of the R-P modes that the 3rd and 4th masses have the largest radius (see inset of Fig.4(b)). Thus, it seems that existence regions of R-P mode are strongly related to the position where the ILM stands.

4.2. Existence regions and stability

For the stability analysis, the characteristics multipliers are computed for each rotating ILM. Since the system of Eq.(1) is a conservative system, ILM is stable if and only if all the characteristic multipliers are located on the unit circle in the complex plane. Otherwise, the rotating ILM is unstable[9].

Figure 5 shows the parameter regions and stability of the ILMs. In the figure, the vertical axis shows the rotation period and horizontal axis shows the initial extension. The colors of the regions correspond to the maximum absolute values of the characteristic multipliers. As shown in the figure, there is the tendency that the range of the rotation period become narrow as the initial extension increases. In addition, the region of the central R-P modes in Fig.3(a) are larger than those of the R-ST modes. The difference may come from the position of ILM and the size of the chain. R-P mode in Fig.3(a) is located exactly at the center of the chain, while another R-ST R-P modes shown in Fig.4(b) is not. If the size of the chain is sufficiently large, the difference of existence region will decrease unless R-P mode stands near the boundaries of the chain. The stability analysis shows that there is a relatively large area, which is colored by blue, where the R-P modes are almost stable. On the other hand, such stable region does not exist for the R-ST modes. Therefore, it is implied that the stability of the rotating ILMs depends on the spatial symmetry of amplitude distribution.

4.3. Existence condition for R-P mode

In the case of the R-P mode, the bifurcation diagram shows that the radius of the rotation approaches zero as the period increases. In the small amplitude regime, Eq.(1) can be linearized. Here we assume the uniform solution that each mass has the same radius of rotation \(r_0\) (see Fig.6).

For the uniform solution, we obtain the following equations from Eq.(1):

\[
I^2 = (1 + a)^2 + (2r_0)^2, \quad (4)
\]

\[
r_0 \left( \frac{2a}{T} \right)^2 = 2 \left[ (l - 1) + 25(l - 1)^3 \right] \frac{2r_0}{T}. \quad (5)
\]

These equations mean that the restoring force of the spring and centrifugal force of angular frequency \(\frac{2a}{T}\) are balanced. By linearizing the equations around \(r_0 = 0\), we obtain a relational expression between the rotation period and initial extension as follows:

\[
T = \pi \sqrt{ \frac{1 + a}{a + 25a^2} }. \quad (6)
\]

The curve in Fig.5(a) is drawn by Eq.(6). As shown in Fig.5(a), the boundary of the existence region of the R-P mode is almost coincide with the theoretical curve. Therefore, it seems that Eq.(6) gives the upper bound of the rotation period of the rotating ILM. Interestingly, Eq.(6) has the same form as Eq.(3) when \(\omega_y = \frac{2a}{T}\).

5. Conclusion

In this paper, we have focused on the rotating ILM in the FPU-\(\beta\) chain and discussed the existence condition and the stability. We have revealed parameter regions where the rotating ILM can exit. As a result, it has been found...
that the region for the R-P mode is wider than that of the R-ST mode. Stability analysis has also shown that the R-P mode has wider stable region than the R-ST mode. In addition, the upper bound of the rotation period has successfully been obtained for the R-P mode which stands at the center of the chain by using a uniform solution and the small amplitude approximation. However, the upper bound is not well-fit to the case of R-ST mode which does not stand on the center. The bifurcation occurs before the amplitude becomes sufficiently small. An analysis without the small amplitude approximation would be required. From the results of the paper, one can expect that a stable or nearly stable rotating ILM exist in a realistic system in which masses are connected nonlinearly, such as the carbon monoatomic chain. In the future, we will search localized solutions in a nonlinear lattice model of the carbon monoatomic chain.

References


Technical Trading Strategy Using Reactions to Stock Price Jumps

Tokimaru Tsuruta and Tomoya Suzuki

Graduate School of Science and Engineering, Ibaraki University
4–12–1, Nakanarusawa-cho, Hitachi-shi, Ibaraki, Japan
Email: {16nm929f, tomoya.suzuki.lab}@vc.ibaraki.ac.jp

Abstract—As a technical indicator, the BPV ratio has been proposed to dynamically detect market price jumps. In the present study, we first confirmed the existence of an anomaly in Japanese and American stock markets where stock price movements are often biased to react on price jumps. Next, we applied this anomaly to decide the dealing position (long or short) by foreseeing this biased reaction of a price movement. Finally, we confirmed the efficiency of this trading strategy through some investment simulations, and also confirmed the validity of the decided dealing positions and their timings by bootstrap statistical tests.

1. Introduction

In financial markets, dealing prices sometimes jump suddenly due to significant news such as natural, political, or financial accidents. According to behavioral economics, market jumps are sometimes caused by overreaction of traders to new information. If so, the following prices might be reversed to the appropriate price by the efficiency of market. To detect these jumps, some techniques[1, 2] have been proposed, and, as one of them, the BPV ratio[3] was proposed as a technical indicator to decide the investment timing. Moreover, our previous study[4] found out that the significant reaction happens right after a market jump detected by the BPV ratio[3], and then showed the validity of the investment strategy applying this significant reaction. However, its validity has been confirmed only in the Japanese stock market. For this reason, in the present study, we investigate the generality of the given results by analyzing real stocks listed on the Japanese Stock Exchange and the American Stock Exchange.

In Sec.3, we analyze 617 Japanese stocks and 1232 American stocks to confirm whether both markets have the anomaly that unnatural reactions happen right after the market jump. Next, as an application of this biased reaction, in Sec.4, we discuss some technical trading strategies based on the market jumps detected by the BPV ratio, and confirm its efficiency through some investment simulations. Finally, in Sec.5, we also perform the statistical significance test to examine the validity of the dealing positions and their timings decided by our trading strategy, comparing to its randomized strategies. In particular, this statistical test is considered quite important to prove the validity of technical analysis, and this approach has been known as the evidence-based technical analysis[5]. Until now, both positive and negative results have been shown in many previous studies[6], and therefore this topic is very controversial and important in the field of quantitative economics.

2. BPV ratio

This section introduces the technical indicator proposed in Ref.[3], which uses the ratio of the realized volatility (RV) and the bipower variation (BPV). First, let us denote the stock price at the time of t as s(t), and then its return rate r(t) can be defined by

\[ r(t) = \log \frac{s(t)}{s(t-1)}. \] (1)

Then, the RV \( v(t) \) is defined by

\[ v(t) = \frac{1}{N} \sum_{a=1}^{N} r^2(t-a+1) \] (2)

and the BPV \( b(t) \) is defined by

\[ b(t) = \frac{1}{N-1} \sum_{a=1}^{N-1} |r(t-a+1)||r(t-a)| \] (3)

where \( N \) is the temporal period of the latest historical data, and we set \( N = 10 \) in the present study. This BPV was proposed by Ref.[1] as a new volatility to detect market jumps, and is composed by two absolute values of temporally adjacent return rates. Therefore, the BPV is less sensitive to large price movements, i.e., market jumps than the RV. Namely, if a market jump occurs, their ratio \( b(t)/v(t) \) becomes smaller. On the other hand, if \( r(t) \) has no jumps or no specific trends during the latest temporal period, that is, \( r(t) \) obeys the normal stochastic process, the ratio \( b(t)/v(t) \) converges into \( 2/\pi \) under the condition of \( N \to \infty \) [1].

For this reason, Ref.[3] proposed the BPV ratio \( c(t) \) so as to converge into one under the condition:

\[ c(t) = \pi \frac{b(t)}{2v(t)} \] (4)

where \( \lim_{N \to \infty} c(t) = 1. \)

However, because \( N \) is a finite value in actual trading, the convergence value is not exactly equal to one. As a solution, we compose the distribution of the latest \( c(t-i) \), \( i = 0 \sim I-1 \), and calculate its mean value \( m(t) \) and standard
deviation $\sigma(t)$. If $c(t-1) > m(t-1)$ and $c(t) < m(t)-k \cdot \sigma(t)$, we consider that a market jump happens at the time $t$ because $c(t)$ becomes smaller than the average level $m(t)$. Here, we set $k = 1$ and $l = 120$ as the same as Ref.[4].

3. Biased Reactions to Market Jumps

According to behavioral economics, market jumps are sometimes caused by overreaction of traders to new information. If so, the following prices might be reversed to the appropriate price. To confirm this possibility, we analyzed 617 Japanese stocks and 1232 American stocks in 2000–2006.

We classified market jumps into two types: the daytime jump and the nighttime jump. The daytime jump means the jump from the opening price to the closing price in the same day, and its reaction is observed at the next opening price. Similarly, the nighttime jump means the jump from the closing price to the opening price, and its reaction is observed until the next closing price. In addition, we classified market jumps into two types: rising jump and falling jump. To detect jumps, we used the BPV ratio.

Table 1 shows the results. We can see that the reversal behavior tends to be happened right after a market jump occurs especially at nighttime. This tendency is almost the same in both the Japanese market and the American market. It might be caused by the over-reaction of traders and its modification into the appropriate price. During the nighttime, because markets are closed, any traders cannot make a deal even if important news happens. In this case, new orders are reserved at night, and all of them are concentratedly dealt with at the next opening time. This over-reaction to news causes an excessive nighttime jump. However, due to the efficiency of the market, this nighttime jump is immediately modified into the proper level. This modification might be a reason why the reversal reaction to the nighttime jump is most clearly shown in Table 1.

4. Technical Trading Strategy with Biased Reactions

As shown in Sec.3, because the stock price tends to move reversely to market jumps (especially to the nighttime jump), we apply this biased reaction to a technical trading strategy. In this section, we prepare the following four strategies for comparison.

(A) We take a short (sell) position at $t$ if $c(t)$ indicates a daytime rising jump.

(B) We take a long (buy) position at $t$ if $c(t)$ indicates a daytime falling jump.

(C) We take a short position at $t$ if $c(t)$ indicates a nighttime rising jump.

(D) We take a long position at $t$ if $c(t)$ indicates a nighttime falling jump.

Table 1: Reaction of the price movement right after a rising or falling jump based on (a) 617 Japanese stocks and (b) 1232 American stocks. Each market jump was observed during the daytime or the nighttime. Bold figure means more than 50% in each category.

(a) Japanese stocks

<table>
<thead>
<tr>
<th></th>
<th>Daytime</th>
<th></th>
<th></th>
<th>Nighttime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rising</td>
<td>Falling</td>
<td>Rising</td>
<td>Falling</td>
</tr>
<tr>
<td>Follow</td>
<td>44.3%</td>
<td>34.9%</td>
<td>35.6%</td>
<td>34.2%</td>
</tr>
<tr>
<td>Do not move</td>
<td>15.2%</td>
<td>15.7%</td>
<td>6.6%</td>
<td>7.7%</td>
</tr>
<tr>
<td>Reverse</td>
<td>40.5%</td>
<td>49.5%</td>
<td>57.7%</td>
<td>58.1%</td>
</tr>
</tbody>
</table>

(b) American stocks

<table>
<thead>
<tr>
<th></th>
<th>Daytime</th>
<th></th>
<th></th>
<th>Nighttime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rising</td>
<td>Falling</td>
<td>Rising</td>
<td>Falling</td>
</tr>
<tr>
<td>Follow</td>
<td>39.5%</td>
<td>37.9%</td>
<td>43.1%</td>
<td>39.1%</td>
</tr>
<tr>
<td>Do not move</td>
<td>17.4%</td>
<td>15.4%</td>
<td>5.3%</td>
<td>5.0%</td>
</tr>
<tr>
<td>Reverse</td>
<td>43.2%</td>
<td>46.7%</td>
<td>51.6%</td>
<td>55.9%</td>
</tr>
</tbody>
</table>

In our strategies (A)–(D), we immediately close the position at the next time $t+1$. According to Table 1, we consider that the strategies (C) and (D) will work well.

Next, as another idea to obtain many trading opportunities, we invest all of the stocks where $c(t)$ indicates a price jump in each time $t$. For example, if price jumps are recognized in three stocks at the same time, we invest these three stocks by the same allocation rate, which is called the dynamically-allocated investment in this study.

To compare the trading performance of four trading strategies (A)–(D), we perform investment simulations using the same kind of stocks analyzed in Sec.3. However, because Sec.3 used the stock data during 2000–2006 to discover the biased reactions, another period should be used for the investment simulation as out-of-sample data. Therefore, we use the following new stock data during 2007–2013 for the investment simulation.

The result is shown in Table 2 and Figure 1. Here, $F$ is the number of days when any trade was executed by the BPV’s signal, and these days are denoted as $t^*$. Then, $R$ is the total return rate [%] during the investment period, and $\bar{R}$ is the average return rate [%] per day, that is, $\bar{R} = R/F$. Moreover, the winning percentage $W$ [%], the payoff ratio $\phi$, and the profit factor $\psi$ are calculated by

$$W = \frac{|\{r^* | r^*(t^* + 1) > 0\}|}{|\{r^* | r^*(t^* + 1) \neq 0\}|},$$

$$\phi = \frac{\langle |R(t^* + 1)R(t^* + 1) > 0\rangle \rangle}{\langle |R(t^* + 1)R(t^* + 1) < 0\rangle \rangle},$$

$$\psi = \frac{\sum(R(t^* + 1)R(t^* + 1) > 0)}{\sum(R(t^* + 1)R(t^* + 1) < 0)} = \phi \cdot \frac{W}{1 - W},$$

where $r^*(t^* + 1)$ is the return of the stock invested at $t^*$, and $R(t^* + 1)$ is our return obtained by closing the position at $t^* + 1$. Then, $\{\}$ means a set, and its number, its average,
Table 2: Investment performance by the strategies (A)–(D) to 617 Japanese stocks and 1232 American stocks. Bold figure means the best score in each category.

<table>
<thead>
<tr>
<th></th>
<th>Japanese stocks</th>
<th>American stocks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A)</td>
<td>(B)</td>
</tr>
<tr>
<td>( F )</td>
<td>1.367</td>
<td>1.272</td>
</tr>
<tr>
<td>( R ) [%]</td>
<td>-19.4</td>
<td>181</td>
</tr>
<tr>
<td>( \bar{R} ) [%]</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>( W ) [%]</td>
<td>4.59</td>
<td>56.9</td>
</tr>
<tr>
<td>( \phi )</td>
<td>1.02</td>
<td>0.99</td>
</tr>
<tr>
<td>( \psi )</td>
<td>0.97</td>
<td>1.37</td>
</tr>
<tr>
<td>( M ) [%]</td>
<td>59.1</td>
<td>19.2</td>
</tr>
</tbody>
</table>

Figure 1: Temporal change of the asset growth rate by the strategies (A)–(D).

and its sum are respectively denoted as \(|\cdot|\), \(\langle\cdot\rangle\), and \(\sum\{\cdot\}\). In addition, we calculate the maximum drawdown rate \( M \), which is the maximum rate of the cumulative losses since the previous largest profit. As expected in Table 1, the strategies (C) and (D) show better investment performance in both markets. Therefore, the biased reactions especially to the nighttime jump are useful to compose technical trading strategies.

5. Statistical Significance Tests for Evidence-based Technical Analysis

Although we could confirm that our strategies, especially (C) and (D), are profitable by using the biased reactions to market jumps, we next confirm their validities by the statistical significance test from the following viewpoints: the investment position of long/short and the investment timing.

5.1. Validity of Investment Positions

First, we investigate whether each trading strategy can take long/short positions properly in the dynamically-allocated investment. For this propose, we randomly shuffle all of the original positions, but keeping each trading timing \( t^* \), and then we evaluate the trading performance by the same measures as Sec.4. For example, if we use \( R \) as a measure, we calculate \( R_r \) \( (r = 1 \sim 10,000) \) for 10,000 randomized strategies shown in Fig.2. Then, we compare them with \( R \) given by the original strategy in the statistical significance test. This method is a sort of nonparametric bootstrap tests.

However, by randomly shuffling the trading positions, the number of each position decided by the randomized strategies becomes different from that for the original strategy. This difference gives either advantage or disadvantage to the randomized strategies. Therefore, we have to modify each investment performance given by the \( r \)th randomized strategy. For example,

\[
R_r \leftarrow R_r + \alpha, \tag{8}
\]

\[
\alpha = \sum_{i=1}^{l} \alpha_i, \tag{9}
\]

\[
\alpha_i = (L_{a,i} - L_{r,i}) \cdot \bar{r}_i - (S_{a,i} - S_{r,i}) \cdot \bar{r}_i, \tag{10}
\]

where \( \bar{r}_i \) is the average return rate of the \( i \)th stock during the investment period, and \( l \) is the total number of stocks \((i.e., i = 1 \sim l)\). Then, \( L_{a,i} \) and \( S_{a,i} \) are the numbers of taking long and short positions of the \( i \)th stock by the original strategy, and \( L_{r,i} \) and \( S_{r,i} \) are those by its \( r \)th randomized strategy. Similarly,

\[
\bar{R}_r \leftarrow \frac{1}{F} [R_r + \alpha], \tag{11}
\]

\[
W_r \leftarrow W_r + \frac{1}{F} \sum_{i=1}^{l} \beta_i, \tag{12}
\]

\[
\beta_i = (L_{a,i} - L_{r,i}) \cdot w_i - (S_{a,i} - S_{r,i}) \cdot w_i, \tag{13}
\]

\[
w_i = \frac{|\{t^* \mid r_i(t^* + 1) > 0\} - \{t^* \mid r_i(t^* + 1) < 0\}|}{\{t^* \mid r_i(t^* + 1) \neq 0\}}. \tag{14}
\]

\[
\phi_r \leftarrow \phi_r \cdot \frac{W_r}{1 - W_r}, \tag{15}
\]

where Eq. (15) is modified by using \( W_r \) given in Eq. (12), but \( \phi_r \) does not need any modification because it is calculated only by the average values in Eq. (6), and therefore the numbers of long and short positions are not essential. Then, it is so hard to modify \( M \) that the maximum draw down is not used in this test.

By comparing the trading performance given by the original strategy to those given by its randomized strategy, we can evaluate the \( p \)-value, that is, the percentage that the randomized strategies can win against the original strategy. If the \( p \)-value is less than 5%, we can conclude that the original strategy is statistically valid. Table 3 shows each \( p \)-value given by the same measure as Table 2. As a result, we can confirm the validity of our strategies (C) and (D) very clearly in terms of almost all measures.

5.2. Validity of Investment Timings

Next, we investigate whether each strategy can decide investment timings properly in the dynamically-allocated
6. Conclusion

In the present study, we applied the BPV ratio to detect market jumps, and confirmed that the reversal reaction of stock price tends to be happened right after nighttime jumps. Then, we proposed a trading strategy based on this tendency, and confirmed its validity through some investment simulations using real stock data listed on the Japanese Stock Exchange and the American Stock Exchange. Moreover, to confirm the validity of our strategy in more detail, we performed the statistical significance tests by comparing it to its randomized strategies. As a result, we could confirm the validity in terms of the investment positions (long or short) and their investment timings. These results are general in financial markets because these were confirmed in two major markets.

This research was partially supported by JSPS KAKENHI Grant Number 16K00320.

References

Koopman Operator Approach to Vital Sign Detection

Yui Kawamura†, Sho Shirasaka†, Hiroya Nakao†† and Wataru Kurebayashi†

†Dept. of Information Engineering, Aomori University
2–3–1 Kohbata, Aomori, Aomori 030-0943, Japan
‡Graduate School of Information Science and Engineering, Tokyo Institute of Technology
‡‡School of Engineering, Tokyo Institute of Technology
2–12–1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan
Email: wkurebayashi@gmail.com

Abstract—Koopman operator approach has attracted much attention in the field of nonlinear time series analysis, because this approach enables us to define nonlinear dynamic modes based on dynamical systems theory only from mixed-mode time series data. Here, we propose a new method for detecting vital signs by the Koopman operator approach, and demonstrate the validity of our method.

1. Introduction

The detection of vital signs, such as heart beats, respiration and body motion, has become a hot topic, because of the recent progress in sensor technology. In these days, we can measure vital signs without touching patients by a novel sensor technology, Doppler sensor [1]. The Doppler sensor utilizes the Doppler effect of microwave signals to measure the motion of a patient’s body including respiratory and cardiac oscillations.

When vital signs are measured by the Doppler sensor, the measured signal is a mixture of several vital signs. Thus, it is necessary to decompose the mixed signal into several kinds of vital signs such as heart beats, respiration and body motion. In clinical applications, signal processing techniques, such as anomaly detection, are used to each vital sign.

In this study, with the aim of establishing a unified framework of the signal decomposition and anomaly detection of vital sign signals, we employ a theoretical approach based on the Koopman operator [2]. By the dynamic mode decomposition (DMD) [3, 4] based on the Koopman operator, we can decompose a mixed signal of vital signs and extract essential dynamical variables, e.g. an oscillation phase of each vital sign. From the Koopman operator reconstructed from measured signals, we can obtain the conditional probability that describes the dynamics of each dynamical variable, which can be used for developing methods of anomaly detection.

2. Signal decomposition and anomaly detection

We introduce a linear operator called the Koopman operator. We assume the existence of a latent stationary Markov process underlying vital sign signals:

\[ x_{t+1} \approx p(x_{t+1}|x_t), \quad (1) \]

where \( x_t \in \Omega \) is the state of the system at time \( t \), and \( p(x|\omega) \) is a conditional probability that describes the dynamics of this system. The Koopman operator \( \mathcal{K} \) corresponding to Eq. (1) is defined as a conditional expectation:

\[ \mathcal{K} f(x) = \int_{\Omega} f(x') p(x'|x) dx', \quad (2) \]

for an arbitrary function of \( x \in \Omega \), \( f(x) \), in an appropriate functional space.

Since the Koopman operator is a linear operator, we can consider the spectral decomposition of this operator, and the eigenvalues and eigenfunctions of the Koopman operator can be computed by the DMD. Let \( \theta_t = \Theta(x_t) \) be an oscillation phase as defined in Ref. [5]. The oscillation phase \( \theta_t \) can be estimated from a mixed signal of vital signs by the DMD, because \( e^{i\lambda(x_t)} \) is an eigenfunction of the Koopman operator. In addition, we can reconstruct a conditional probability \( q(\theta_{t+1}|\theta_t) \) from time series data. From this conditional probability, it is expected that anomalies in the data can be detected by thresholding, because such rare events have much lower transition probability than the normal dynamics. In the presentation, we will demonstrate the validity of our method.

References

Large Graph Laplacian Matrix and Functional Map of Whole Brain of C. elegans

Hiromu Sakuma †, Takayuki Teramoto ‡, Sayuri Kuge ‡, Takeshi Ishihara ‡ and Yuishi Iwasaki †

†Graduate School of Science and Engineering, Ibaraki University
4–12–1 Nakanarusawa, Hitachi, Ibaraki 316-8511, Japan
‡Graduate School of Science, Kyushu University
744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
kuge.sayuri.449@m.kyushu-u.ac.jp, takeiscb@kyushu-u.org,
yuishi.iwasaki.sc@vc.ibaraki.ac.jp

Abstract—Towards understanding of neural signaling of Caenorhabditis elegans, cluster analysis is carried out for the central whole brain imaging data on the basis of spectral clustering. Correlation between the data is represented as weighted edges in a similarity graph. Optimized clustering resolves into an eigenvalue problem of the graph Laplacian defined by the similarity graph. We analyze the neural activities of the wild-type and the unc-7 mutant which has defect in gap junction. In the wild-type, the neural activities within the same cluster are highly coherent. There are anti-phase clusters in which the neural activities are obviously in anti-phase. In the unc-7 mutant, highly coherent neural activities are disappeared. Gap junction is required to generate a highly organized neural synchronization. In functional maps of the neurons, the functional left-right symmetry of neuronal position within the same cluster is partially shown.

1. Introduction

The nematode Caenorhabditis elegans is subject to the brain activity map project together with Drosophila, zebrafish and mouse [1]. C. elegans is a useful model organism in neurobiology because the neural connectivity is fully known [2]. The nervous system of C. elegans consists of 302 neurons connected by about 6500 chemical synapses and about 900 gap junctions. About 170 neurons of them densely locate in the head region. Although the wiring diagram of the nervous system is determined, functional map is poorly known. Recently, we developed a 4D imaging system to measure the neural activity in the head region as a worm lives [3]. Therefore, cluster analysis is carried out for the central whole brain imaging data.

The k-means algorithm and the hierarchical clustering are popular clustering methods. Since the k-means algorithm has a tendency to divide objects into equally sized clusters, an obviously inappropriate partition is frequently occurred. In the hierarchical clustering, objects tend to be added to the tail of the largest cluster. This is known as “chaining phenomenon” and often generates spurious clusters. In this work, spectral clustering [4] is applied to determine appropriate clusters of the neural imaging data. Correlation between the data is represented as weighted edges in a similarity graph. The graph Laplacian is subsequently defined by the similarity graph. Spectral clustering finds the eigenspace of a given data set on the basis of the eigenvectors of the graph Laplacian.

2. Method

2.1. Graph Laplacian

For a given graph with N vertices, its $N \times N$ graph Laplacian $L$ is defined by

$$L = D - W,$$  (1)

where $D$ and $W$ are the $N \times N$ degree matrix and the $N \times N$ adjacency matrix, respectively. The element $w_{ij}$ of $W$ corresponds to an edge for a directed graph or an arc for an undirected graph whose value represents the weight from $i$-th vertex to $j$-th vertex. The element of $D$ is $d_{ii} = \sum_{j=1}^{N} w_{ij}$ or $d_{ii} = 0$ for $i \neq j$.

There are several important properties in the graph Laplacian. Let $\lambda_k$ and $\xi_k^i$ be the eigenvalue and the eigenvector of $L$, respectively.

$$L \xi_k^i = \lambda_k \xi_k^i.$$  (2)

When $w_{ij}$ is non-negative and symmetric, that is $w_{ij} = w_{ji} \geq 0$, the eigenvalues are real and non-negative. The smallest eigenvalue is 0 and its eigenvector is $\vec{1} = (1, 1, \cdots, 1)$. When the eigenvalues are arranged in an ascending order, therefore, $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$. The multiplicity $M$ of $\lambda_k$ is $0 (k = 1, \cdots, M)$ is equal to the number of connected components $G_1, \cdots, G_m$ in the graph. The eigenspace of $\lambda_k = 0$ is spanned by the vectors $\vec{1}_{G_1}, \cdots, \vec{1}_{G_m}$ of those components. Spectral clustering is indebted to these properties.
2.2. Spectral clustering

First of all, a similarity graph in which each vertex represents each data is considered. A certain degree of “similarity” between the data, whose value \( w_{ij} \) is non-negative and symmetric, is attached as an edge so that the similarity graph is undirected. In this work, the similarity graph is divided into \( K \) clusters \( G_1, \ldots, G_K \) whose partition satisfies \( G_i \cap G_{i'} = 0 \) for \( i \neq i' \) and minimizes the following cost function [5].

\[
\text{NCut}(G_1, \ldots, G_K) = \sum_{k=1}^{K} \frac{\text{cut}(G_k, \bar{G}_k)}{\text{vol}(G_k)}
\]

(3)

Where \( \text{cut}(A, B) = \sum_{i \in A, j \in B} w_{ij}/2 \), \( \bar{A} \) is the complement of \( A \) and \( \text{vol}(G_k) = \sum_{u \in G_k} d_u \). The number of clusters \( K \) is a given number. The cost function in eq. (3) indicates the sum of the edges \( w_{ij} \) between \( G_k \) and \( \bar{G}_k \) normalized by the factor \( \text{vol}(G_k) \).

Since this optimization problem is NP-hard [6], it is difficult to determine optimal clustering for a large graph. In the case of nervous system of \( C. elegans \), for example, the number of vertices is \( N \approx 170 \) for the head region or \( N = 302 \) for the entire nervous system. According to the Rayleigh-Ritz theorem, the cost function in eq. (3) is approximately replaced with the sum of the eigenvalues \( \lambda_k' \) of the normalized graph Laplacian \( L_{\text{sym}} = D^{-1/2}LD^{-1/2} \) for optimization.

\[
\min_{G_1, \ldots, G_K} \text{NCut}(G_1, \ldots, G_K) \rightarrow \min_{x_1, \ldots, x_K} \sum_{k=1}^{K} \lambda_k'.
\]

(4)

Practically, the optimization problem resolves into an eigenvalue problem of \( L_{\text{sym}} \). Spectral clustering algorithm is as follows [4].

**step 1.** First of all, \( W \) of the similarity graph is given. Subsequently, \( D, L \) and \( L_{\text{sym}} \) are defined from \( W \).

**step 2.** The eigenvalues \( \lambda_k' \) and the eigenvectors \( \overrightarrow{\xi}_k \) of \( L_{\text{sym}} \) are calculated: \( L_{\text{sym}} \overrightarrow{\xi}_k = \lambda_k' \overrightarrow{\xi}_k \).

**step 3.** Using \( \overrightarrow{\xi}_k \) corresponding to the smallest \( K \) eigenvalues, \( N \times K \) matrix \( U = (u_{ik}) \) is introduced. Where \( u_{ik} = \xi_{ij}'/\sqrt{\sum_{i}^{N} (\xi_{ij}')^2} \). Let \( \overrightarrow{v}_i \) be the \( K \) dimensional vector corresponding to \( i \)-th row of \( U \). The points \((v_i)_{i=1, \ldots, N} \) are divided into \( K \) clusters \( C_1, \ldots, C_K \) by the \( k \)-means algorithm.

In this work, \( W \) is given by the Pearson’s correlation coefficient \( r_{ij} \) between \( i \)-th and \( j \)-th neuron data. Since \( w_{ij} \) must be non-negative in a similarity graph, the absolute value of \( r_{ij} \) is introduced to define the graph Laplacian, that is \( w_{ij} = |r_{ij}| \). A negative correlation works in the same way as a positive correlation in the spectral clustering. For examples, the data in Fig.1a, Fig.1b and Fig.1c probably belong to the same cluster although the correlation coefficient \( r_{bc} = -0.662 \) is large negative. The sign of \( r_{ij} \) is not taken into account.

All programs in this work are written in the C language and are compiled by the GNU C compiler on UNIX.

2.3. Experimental data

For the neurons in the head region of \( C. elegans \), somatic \( \text{Ca}^{2+} \) concentrations are simultaneously visualized by the 4D live-cell imaging system [3]. Voltage-gated \( \text{Na}^+ \) channels have not been found in \( C. elegans \) [7]. Instead of \( \text{Na}^+ \)-based classical action potentials, the neurons might have \( \text{Ca}^{2+} \)-based signal amplification as in the large nematode Ascaris. Therefore, the \( \text{Ca}^{2+} \) signaling well reflects the dynamics of the neural activity. Using the \( \text{Ca}^{2+} \)-binding fluorescent protein YC2.6, \( \text{Ca}^{2+} \) imaging data is measured for the wild-type (the standard “normal” type) and the \text{unc}-7 mutant. The \text{unc}-7 mutant has defect in gap junction and exhibits uncoordinated locomotion. The neural connectivity of the \text{unc}-7 mutant is different from that of the wild-type. No stimulation is added to a worm during measurement. Time interval is 1/4.5 sec. In this work, we adopt the time course data sets of 2000 time points, \( x_{i}(n) \) \((n = 1, 2, \ldots, 2000) \). Here \( i \) and \( n \) are neuron and time indices, respectively. The number of observed neurons is about \( N \approx 150 \). To eliminate the YC2.6 expression difference in each neuron, we analyze the normalized values \( (x_{i}(n) - x_{\text{min},i})/(x_{\text{max},i} - x_{\text{min},i}) \). \( x_{\text{max},i} \) and \( x_{\text{min},i} \) are the maximum and the minimum values of \( x_{i}(n) \), respectively. Examples of the neural activity are shown in Fig.1.
3. Results

3.1. Spectrum of graph Laplacian

Spectra of the normalized graph Laplacian $L_{sym}$ are shown in Fig. 2. We confirm that the smallest eigenvalue of the graph Laplacian is $\lambda_1 = 0$. Eq. (4) indicates that the $(K + 1)$-th eigenvalue $\lambda_{K+1}$ gives the sum of newly cut weights $w_{ij}$, that is an additional "cost", when the number of partitions is increased from $K$ to $(K + 1)$. At a large spectral gap, therefore, no loosely connected cluster remains. The graph is already divided into strongly connected clusters. Since a large spectral gap is not shown in Fig. 2, an appropriate $K$ is not determined to indicate a distinct division of the graph. The graph is connected via the edge $w_{ij}$ which takes a continuous value between 0 and 1. Therefore, the spectra of the graph Laplacian become continuous.

When $W$ is binarized using the threshold $\varepsilon$ such as $w_{ij} = 1$ for strong correlation $|r_{ij}| \geq \varepsilon$ or $w_{ij} = 0$ for weak correlation $|r_{ij}| < \varepsilon$, the spectrum of the graph Laplacian has a large gap. When $\varepsilon = 0.4$, for an example, the graph Laplacian has a spectral gap at $k = 4 \sim 7$ (results not shown in this paper).

3.2. Functional map of neurons

In the case of $K = 4$, the spectral clustering of the central whole brain imaging data are shown in Fig. 3. In the wild-type (Fig. 3a), the neural activities within the same cluster are highly coherent. The clusters $C_2$ and $C_3$ are in anti-phase. The clusters $C_2$ and $C_4$ are almost in phase. There are many connections among $C_1$, $C_2$ and $C_3$. On the other hand, $C_4$ has a few connections between the other clusters. As the number of partitions $K$ increases, the clusters are divided into small clusters in which the neural activities are coherent. Correlation between the clusters are consistent with that of the synaptic connectivity in $C$. elegans (results not shown in this paper).

In the unc-7 mutant (Fig. 3b), highly coherent neural activities are disappeared. The clusters $C_2$ and $C_4$ are in anti-phase. The cluster $C_3$ has a few connections between the other clusters. In addition to the unc-7 mutant, we analyze the unc-13 mutant which has defect in chemical synapse. In the unc-13 mutant, highly coherent neural activities are also disappeared (results not shown in this paper). We find that both gap junction and chemical synapse are required to generate a highly organized neural synchronization.

Functional maps of the neurons are shown in Fig. 4. The neurons in the same cluster are filled with the same color and are located at the measurement point in a worm. The neural activity is measured in soma. Each neuron extends long axons from its soma and is connected to other neurons at end point of the axon. Therefore, a physical distance between the somas is not related to a "similarity distance" between the neural activities in the somas. In Fig. 4, the neurons in the same cluster are not localized spatially. In
4. Discussion

Spectral clustering using the normalized graph Laplacian is applied to divide the neural activities in C. elegans. Instead of eq. (3), the following cost function is considered [8].

\[
\text{RatioCut}(G_1, \cdots, G_K) = \sum_{k=1}^{K} \frac{\text{cut}(G_k, G_{\bar{k}})}{|G_k|},
\]

where |A| is the number of vertices in A. The spectral clustering with eq. (5) is achieved by solving the eigenvalue problem of the graph Laplacian \(L \in \mathbb{R}^{n \times n}\) in eq. (2). In this case, spectral clustering divides the neurons into a large cluster and other tiny clusters for the wild-type and the \(unc-7\) mutant. As the number of partitions \(K\) is increased, an additionally divided cluster is tiny.

To provide a large spectral gap of the graph Laplacian, \(W\) of the similarity graph must be sparse. \(W\) represents a neighborhood graph in which the “thin” edges are removed. Although the correlation coefficient is used to determine \(W\) in its work, binarized graph Laplacian \(L\) is able to be determined directly from the neural data [9].

Since individual difference in the neuron distribution is quite large, neuron annotation, that is a one-to-one correspondence between the imaging data and neuron name, is poorly succeed. When the annotation is completed, the presented cluster analysis is helpful to understand neural signaling of C. elegans.

Acknowledgments

The authors would like to thank Y. Iino, Y. Toyoshima, T. Tokunaga, O. Hirose and R. Yoshida for valuable discussions on \(Ca^{2+}\) imaging, and K. Takeda for helpful suggestions on spectral clustering. This work was supported by CREST program of Japan Science and Technology Agency.

References


A Construction method for disjoint paths with transmission reliability on communication networks

Sho Nishi †, Takayuki Kimura ‡ and Kenya Jin’ no †

†Graduate School of Electronics, Information and Media Engineering, Nippon Institute of Technology
‡Department of Electrical and Electronics Engineering, Nippon Institute of Technology

Email: †2158013@estu.nit.ac.jp, ‡tkimura,jinno}@nit.ac.jp

Abstract—The transmission networks such as mobile communication networks continues to expand because of the huge growth of users. Even if the networks increase, the Quality of Service should be maintained. To overcome this problem, finding the working paths and the efficient alternative paths are used to guarantee transmissions on the communication networks. A conventional technique constructs both a working path and an alternative path for transmission, however, the information of the other paths are ignored. In this study, we propose a construction method for the working paths and its alternative paths using the information from the other paths, achieving our proposed method removes the packet congestion in comparison to the conventional techniques.

1. Introduction

Recently, the QoS guarantee degrades because of link (or server) failures in the network by the expansion of the transmission network. If the transmission between users is not established suitably, these users frustrate because of the failure of access. An important key to overcome this problem is reliability of the communication network. The reliability of the communication network represents a probability of being able to establish communication path on the communication network even if failures in the network have occurred. To tackle this problem, two types of transmission paths are prepared before transmission of a packet. One is a working path used for common transmission and another is an alternative path used if the working path fails. The alternative path is constructed using the node set without used by its working path. In general, two paths are denoted as “disjoint paths.” If we find the working path and the alternative path whose reliability are high, the QoS in the communication networks increases. To construct these efficient paths, the working path and the alternative path should be disjointed; the node and link set used by the working paths and those by alternative paths are independent. Finding the best combination of disjoint paths is the non-linear optimization problem. Further, because the communication networks have the scale-free property [2], the node distribution by constructing the disjoint paths is biased. Figure 1 shows how many times each node is used for the disjoint paths. In these simulations, we set the number of nodes of network to 100, and cost of each link to 1. In Fig. 1, many paths go through the nodes whose degree are larger than the others. The congestion of packets occurs on these nodes. The QoS is then degraded by the congestion on the communication network. Therefore, finding the effective disjoint paths is important not only to enhance the reliability of packet transmission but also to avoid the congestion in the network.

In this study, we propose a new method for finding the effective disjoint paths using availability of nodes in the networks. To evaluate the effectiveness of our the proposed method, we use the probability graph whose links fail or broken with a probability. Results of simulations indicate that our proposed method constructs better disjoint paths, achieving the congestion is well removed.

![Figure 1: Node distribution of the disjoint paths by the conventional method for the scale-free networks.](image)

2. Problem formulation

We use an undirected network $G = (V, L)$, where $V = \{v_1, v_2, v_3, \ldots, v_N\}$ is a set of nodes and $L$ is a set of links as a network model. Each link $(i, j) \in L$ has a failure probability...
A cost of each link, $c_{ij}$, is then defined as follows:

$$
c_{ij} = \begin{cases} 
\lambda & \text{if } (i, j) \in L, \\
\infty & \text{otherwise}.
\end{cases} \quad (1)
$$

A path from node $s$ to node $t$ in the network is described by $p_{st} = \{s, (s, i), i, \ldots, j, (j, t), t\}$. A cost of a path $p_{st} = \{s, (s, i), i, \ldots, j, (j, t), t\}$ is then defined as follows:

$$
C(p) = \sum_{i \neq ji} b_{ij} + \sum_{(i, j) \in L} c_{ij}.
$$

(2)

where $b_{ij}$ is the number of times the $i$th node is used for the working paths, $c_{total}$ is the total costs in the network. A transmission reliability of a path is then defined as follows:

$$
Tr(p_{st}) = e^{-C(p_{st})}.
$$

(3)

The most reliable pair of disjoint paths from $s$ to $t$ is given by,

$$
\text{max} Tr(p_{st}^w \cup p_{st}^a) = Tr(p_{st}^w) + (1 - Tr(p_{st}^w))Tr(p_{st}^a),
$$

(4)

where $p_{st}^w$ is a working path from $s$ to $t$, $p_{st}^a$ is an alternative path from $s$ to $t$. The objective function for finding the effective disjoint paths is then defined as follows:

$$
\text{max} \ Tr(G) = \sum_{w, a} \text{max} \ Tr(p_{st}^w \cup p_{st}^a),
$$

(5)

where $G$ is an undirected network.

3. Path searching algorithm

Although the conventional method for the disjoint paths [1] uses predefined link costs for constructing the working and alternative paths, we employ the optimum path strategy [4] which considers the node betweenness centrality to assign the weight of the link. By using the optimum path strategy [4], we expect that the biased node distribution for making the paths is equalized and the congestion of networks is alleviated. An algorithm of our proposed method in this study is described as follows:

-step 1 Searching the shortest paths between all of the nodes.

$$
p^w \leftarrow \text{Shortest Path Search}(V, L, W(t)),
$$

where $V$ is the set of nodes, $L$ is the set of links, $W(t)$ is the set of weight of link at the $t$th time. Here, $w_{ij}(t)$ is the weight of the links between the $i$th node and the $j$th node at the $t$th time. We set $w_{ij}(0)$ to $c_{ij}$ in Eq. (1).

-step 2 Adding the weight to the link.

To avoid the nodes used by the working paths for constructing alternative paths, we add an weight to links to which the node whose betweenness centrality is the highest among the whole working paths is connected. The adding weight of link is defined as follows:

$$
w_{ij}(t + 1) = \beta w_{ij}(t),
$$

where $\beta$ is a controlling parameter of adding link cost. In this study, we set $\beta$ to 2.

-step 3 Searching the alternative paths by using the weight $W(t + 1)$.

$$
p^a \leftarrow \text{Shortest Path Search}(V, L, W(t + 1)),
$$

-step 4 Calculating the transmission reliability.

The transmission reliability of the network is calculated using the Eqs. (2) – (5).

We repeated from the step 1 to the step 4 until a terminating condition is satisfied. In this study, we repeated these steps 50 times.

4. Simulation

To investigate the reliability of our proposed method, we use a probabilistic graph in these simulations. In the probabilistic graph, each link have failure probability. In these simulations, the following conditions are assigned to the networks.

- A state of link is either normal or failed.
- Each link has an independent probability for link failure.
- States of all links are determined using the link failure probabilities at each iteration.
- States of failure links are repaired after a certain period of time.

In these simulations, we evaluated the congestion rate [5] for the conventional method [1] and the proposed method. To do this, we generated the packets on the networks. We then changed the states of links using the probabilistic graphs and evaluated the congestion rate. We use two types of network. The first is the scale-free network [2] (Fig. 2) and another is the random network [3] (Fig. 3). The number of nodes in the network, $N$, is set to 100. In these simulations, each node has a buffer to store the packets. The packets in the buffer is transmitted to the adjacent nodes using First-In First-Out(FIFO) principle. In addition, each node transmits a packet to its adjacent node at each iteration. A single iteration is defined by the transmission of the packets to the nodes. We generate $R$ packets in the network at each iteration. The source node and the terminal node of each packet are randomly determined. A packet is
removed if the node to which the packet has been transmitted has no empty in its buffer. In addition, a packet is also removed if the packet goes through the failure link. Further, if a packet arrives at its terminal node, the packet is removed. A packet is first transmitted to its terminal node using the working path. If the packet is removed from the network, the packet is again transmitted to its terminal node using the alternative path. In this study, the failure probability of each link is set to 1% and a link is recovered after five iterations. The buffer size of the node is set to 300. We set the number of iterations $T$ to 1,000. Next, the network congestion rate, $0 \leq \eta \leq 1$, is defined as follows:

$$\eta = \frac{\sum_{t=1}^{T} (n(t) - n(t-1)) }{RT},$$

where $n(t)$ is the number of packets in the network at the $t$th time. If $\eta$ is small, the network is free flow, otherwise, the network is congested. The retransmission rate, $0 \leq \delta \leq 1$, is defined as follows:

$$\delta = \frac{x}{U},$$

where $x$ is the number of retransmissions, $U$ is the number of transmissions. If $\delta$ is small, the packet transmissions are successfully conducted, otherwise, the transmissions are frequently interrupted by buffer constraint or link failure. Figure 4 shows the number of times the nodes are used by the disjoint paths by our proposed method. In Fig 4, the nodes with large degree show less employed rate than the ones by the conventional method (Fig. 1) to construct the disjoint paths. The congestion rates and the retransmission rates for the scale-free networks by our proposed method and the conventional method are shown in Figs. 5 and 6. In Figs.5 and 6, the proposed method shows lower congestion rate than the conventional method if the number of generating packets at each node is between 5 and 25. The congestion rate and the retransmission rate for the random network by both methods are shown in Figs. 7 and 8. In Figs. 7 and 8, the proposed method shows lower congestion rate than the conventional method if the number of generating packets at each node is between 5 and 25. From these results, our proposed method shows better congestion rate for the scale-free network than those by the random network. This is because the hub nodes in the scale-free networks are avoided for constructing disjoint paths.
Figure 6: Retransmission rate for the scale-free network

Figure 7: Congestion rate for the random network

Figure 8: Retransmission rate for the random network

5. Conclusions

In this study, we proposed a new construction method for disjoint paths that effectively avoids the congestion at the hub nodes in the communication networks. We evaluated our proposed method by the congestion rate and the retransmission rate. From the results of numerical simulations, our proposed method shows lower congestion rate and retransmission rate than the conventional method. In future works, we evaluate our method by various realistic network conditions to investigate an applicable possibility of our method to the real world system.

The research of T. K. was partially supported by a Grant-in-Aid for Young Scientists (B) from JSPS (No. 16K21327).

References


Efficient Board Feature Extraction for Strategy Improvement in Computer Go

Hayato Mitsuoka and Koujin Takeda

Department of Intelligent Systems Engineering, Ibaraki University
4-12-1, Nakararasawa Hitachi, 316-8511 Ibaraki, Japan
Email: 15nm925f@vc.ibaraki.ac.jp, koujin.takeda.kt@vc.ibaraki.ac.jp

Abstract—Recently, significant progress was made in computer Go by deep learning. However, huge computer resource is required for achieving professional player’s skill at present, which seems over-engineering for feature extraction on Go board. From this background, we discuss how to construct an efficient feature extraction method on Go board under deep learning framework. By making use of knowledge on image recognition by deep learning, we propose a method to reduce computational cost of board feature extraction without degrading Go playing performance.

1. Introduction

In artificial intelligence study, several famous board games have been intensively investigated over years, because the conquest of them by artificial intelligence is thought to be a milestone. Regarding chess or Shogi, computer playing programs have already been able to compete or beaten most proficient professional players.

Go is known as the most difficult board game for artificial intelligence, due to a huge number of possible combination of moves and the equivalence of each stone value unlike chess or Shogi. Nevertheless, steady progress in computer Go program has been made for decades. Monte Carlo tree search[1, 2] is one of the important ideas for the progress, which enables to find an appropriate next move by pruning unnecessary branch of moves on the game tree after self-random play by computer. Even with such progress, the best computer Go program could not compete with any professional Go player at all, at the stage of several years ago.

In this year, quite significant breakthrough was made by AlphaGo[3] by Google DeepMind. Their success was achieved with deep learning. Incorporation of deep learning into Go programming has also been attempted formerly by other groups[4, 5, 6] (a pioneering work with this strategy is found in [7]), and in AlphaGo they also implemented very deep neural network with a huge number of neurons like other studies. Then their program finally beat a professional player. Furthermore, by self-reinforcement learning, AlphaGo surprisingly won against the most outstanding Go player in the world, which means that the technology of the best computer Go program at last surpasses the human player strategy.

However, in AlphaGo, there remains computational cost problem. In AlphaGo they used deep convolutional neural network (DCNN), in which there are about millions of neurons on layers and kernels between layers. For the operation of their DCNN, thousands of CPUs and hundreds of GPUs are required in the whole hardware, which cannot be prepared in personal use. In DeepMind’s project, their final objective of deep neural network study is the realization of artificial intelligence for general purpose use, and their hardware specification may be over-engineering for computer Go.

In this article, we revisit the structure of deep neural network in computer Go, and discuss how to simplify it without degrading Go playing performance. As stated, DCNN is implemented in AlphaGo program, which is in general used mainly in image recognition. This fact suggests that the “picture” of black/white stone pattern plays an important role for human’s decision of next move. In image recognition with DCNN, various ideas are proposed for reduction of computational cost. DropConnect[8] is one of such ideas, where a fraction of edges are disconnected. Accordingly, it enables us to reduce the computational cost in learning and to avoid overfitting problem. We apply DropConnect to computer Go program, and discuss how it affects Go playing performance by the experiment of next move prediction. If DropConnect does not highly degrade the performance of next move prediction, we will be able to reduce the computational cost of DCNN without paying high price. As a consequence, we expect to obtain a key to the method of efficient board feature extraction by DCNN.

2. Model

DCNN is one of multi-layered neural networks, which is mainly used for image recognition. DCNN consists of multiple convolutional/pooling layers, which are alternately layered. Kernel filters are defined on edges between layers. By the operation of convolution as in Figure 1, the system naturally handles overlapped information, which enables us to cope with defect or displacement in original image appropriately and leads to more successful image recognition.

DCNN is also used for board feature extraction on Go board. In computer Go, basic input information is stone pattern on board, and additional set of tactical board information is also often taken as input. The objective of board feature extraction is the correct next move predic-
Figure 1: An example of convolutional operation between layers by kernel filters: In this example, the size of input information is $8 \times 8$. The size of kernel filter is $3 \times 3$. By operation of convolution, output information on the next layer is $6 \times 6$.

Figure 2: Our DCNN architecture: We show 4-layer DCNN. We input 3 channels with $11 \times 11$ size including boundary; black/white stone pattern and board boundary. The size of neuron gradually decreases as $9 \times 9$ (1st) $\rightarrow$ $7 \times 7$ (2nd) $\rightarrow$ $5 \times 5$ (3rd) $\rightarrow$ $3 \times 3$ (4th) by the operation of convolution. The network between final intermediate layer and output layer is fully connected network of $1 \times 1$ size neurons. The output layer has 81 neurons for next move prediction on $9 \times 9$ board.

Then we propagate input information to the neuron on the next layer by the operation of kernel filter and activation function. We use DCNN with 2 or 4 intermediate layers, where all layers are convolutional layers. We do not use pooling layer for feature extraction, as pooling layer is not included in DCNN in preceding works of computer Go. On each intermediate layer 16 neurons are located. We always use $3 \times 3$ size kernel filter. Accordingly, on final intermediate layer we put 16 neurons with $7 \times 7$ size for 2-layer DCNN, and with $3 \times 3$ size for 4-layer DCNN. Finally, for propagation to output layer we divide these neurons into pieces of $1 \times 1$ size, and make fully-connected network with output layer. Output layer consists of $81 (=9 \times 9)$ neurons of $1 \times 1$ size, which describes the position on board. The neuron with the largest value is taken as the result of next move prediction.

3. Method

3.1. Architecture of DCNN in our experiment

Even if we simplify the structure of neural network, training of DCNN requires much time for standard $19 \times 19$ board. Therefore, in this work we restrict ourselves to the case on small $9 \times 9$ board.

The network structure in our framework is shown in Figure 2. For a single channel input, we prepare $11 \times 11$ size image for $9 \times 9$ go board including boundary of width one. We prepare 3 input channels totally; for black stone pattern, white stone pattern, and the information of board boundary.
possible by training. We follow the scheme of standard supervised learning on multi-layer neural networks. We prepare stone pattern on board in actual game as input, and corresponding next move as output. For training we must determine a loss function to minimize, and we choose mean squared error as the loss function $E$.

$$E := \sum_n E_n = \sum_n \left( \sum_k \frac{1}{2} (y_{nk} - t_{nk})^2 \right). \quad (1)$$

In the above definition subscript $n$ represents the label of training data, and $k$ the label of output neuron. $y_{nk}$ is the output signal from DCNN and $t_{nk}$ is the correct signal of training data, which indicates the correct next move. $E_n$ is the loss function for data with label $n$. As the activation function $h(x)$ for forward propagation, we use rectified linear function,

$$h(x) := \max\{0, x\}. \quad (2)$$

After having output $y_{nk}$ by forward propagation, we change values of kernel filter entries by standard back propagation algorithm for minimization of $E$. Here we apply stochastic gradient descent method with Hadamard product (AdaGrad) for back propagation.

$$g_{ij}^{(r+1)} = g_{ij}^{(r)} + \frac{E_n}{\partial W_{ij}^{(r)}}^2,$$

$$W_{ij}^{(r+1)} = W_{ij}^{(r)} - \frac{\alpha}{\sqrt{g_{ij}^{(r+1)}}} \frac{E_n}{\partial W_{ij}^{(r)}}. \quad (3)$$

$W_{ij}^{(r)}$ is the weight of kernel filter between neuron $i$ on an intermediate layer and neuron $j$ on the next layer at $r$th iteration. Note that $W_{ij}^{(r)}$ is the matrix(kernel filter) and the subscript of matrix element is omitted. $g_{ij}^{(r)}$ is an auxiliary variable at $r$th step for computation of $W_{ij}^{(r)}$. $\alpha$ is learning rate and $\alpha = 5 \times 10^{-3}$ in our experiment.

In equation (3), we must calculate the derivative of $E_n/\partial W_{ij}$ analytically, as $E_n$ is expressed as a function of $W_{ij}$. This can be done by standard back-propagation formulation.

### 3.3. DropConnect

DropConnect[8] is a method of training by regularizing neural network with a huge number of edges. In DropConnect, we randomly select a fraction of edges and set them zero, namely we disconnect them. By DropConnect, reduction of overfitting is expected in general, when we remove an appropriate fraction of edges. As a result, we can reduce the computational cost of forward/backward propagation.

Our key idea of this work is; we apply this method to move prediction in Go to reduce computational cost and to avoid overfitting.

![Figure 3: Dependence of next move prediction accuracy on the ratio of disconnection.](image-url)

### 4. Experiment

#### 4.1. Programming environment and dataset

For programming of DCNN we use tiny-cnn[9], which is a C++ library for deep learning. As training/test dataset for next move prediction experiment on $9 \times 9$ board, we use "Igo Quest"[10] dataset by 50 best rating players among all. We assign 7500 games (including 362508 board patterns) to the training data, and 116 games (including 51616 patterns) to the test data for evaluation of next move prediction accuracy.

#### 4.2. Move prediction accuracy (1): dependence on DropConnect and intermediate layer depth

We evaluate next move prediction accuracy under DropConnect by numerical experiment. In DropConnect, we remove a fraction of edges between 1st and 2nd intermediate layers (totally $16 \times 16 = 256$ edges), or equivalently kernel filters on corresponding edges are set to be null matrix. We vary the ratio of disconnection from 0 to 0.9 by 0.1 step. We also vary intermediate layer depth: DCNNs with 2 and 4 intermediate layers are used. Before evaluation of accuracy, we train DCNN with 30 training iterations.

The result is shown in Figure 3. First, the accuracy is almost constant with respect to the ratio of disconnection, and does not show clear fall-off both for 2- and 4-layer DCNNs. Even if we disconnect 90% of edges, the accuracy is almost the same as the fully-connected case. This indicates that the edges in the original fully-connected DCNN is redundant for $9 \times 9$ board feature extraction, many of which can be removed without degrading move prediction accuracy.

As for intermediate layer depth, the accuracy by 2-layer DCNN slightly surpasses 4-layer DCNN. This suggests that training by 4-layer DCNN may lead to slight overfitting, and yield poorer performance than 2-layer DCNN.
Therefore, under our experimental setup, we conclude that 2-layer DCNN is sufficient for next move prediction. However, we should note that the value of accuracy by our experiment is much smaller than by AlphaGo. For next move prediction, AlphaGo yields accuracy of 0.57 for standard 19 × 19 board. Hence, we should keep in mind that DropConnect may worsen move prediction performance under different experimental setup, for example higher accuracy case.

4.3. Move prediction accuracy (2): dependence on training duration

We also study the dependence of accuracy on training duration under 2- and 4-layer DCNNs. The result is depicted in Figure 4, which shows the accuracy by 2-layer DCNN increases more rapidly and saturates earlier than 4-layer DCNN. From this result, we conclude that 2-layer DCNN is more suitable than 4-layer DCNN for fast training, which also helps the reduction of huge training time for DCNN.

We also vary the ratio of disconnection in this experiment, which do not affect the speed of training as a consequence.

5. Summary and Discussion

We discussed how to reduce the computational cost for training and move prediction by DCNN. Our result indicates that we can reduce the number of intermediate layers and edges (=kernel filters) without degrading performance of next move prediction, which leads to the reduction of computational cost.

As future works, we should verify how additional board information inputs affect the result. Furthermore, we should also study the dependence on learning algorithm: Improvement of move prediction accuracy might be possible when we replace the loss function with more appropriate one, or change AdaGrad to another gradient descent algorithm.

Experiment on standard 19 × 19 board is of course a remaining work. After having the result on standard board, we will combine our method with Monte Carlo tree search algorithm to verify playing performance in actual Go games.

Acknowledgments

This work is supported by KAKENHI Nos. 24700007, 25120013 (KT).

References

Internet traffic anomalies and their detection techniques

Kensuke Fukuda

NII / Sokendai
2-1-2 Hitotsubashi, Chiyoda, 101-8430, Japan
Email: kensuke@nii.ac.jp

The Internet has been a crucial infrastructure in our daily life. Most of traffic (packets) are legitimate one, however, we frequently observe some anomalous traffic hidden in such majority of legitimate traffic. These traffic are due to malicious activities such as virus and DoS (Denial of Service) or misuse of devices (e.g., router configuration).

It is not easy task to find such anomalous behavior in Internet backbone traffic. There are two approach to find anomalies in Internet traffic; signature-based and statistics-based one.

The former assumes on an exact matching of packet field to pre-defined anomalous signature similar to recent virus detection techniques. The accuracy of the detection is high, however the drawback of this approach is a difficulty to scale up to current high speed networks (e.g., 10-100Gbps networks). Furthermore, it naturally requires known packet signatures, thus it is impossible to find recent so-called zero-day attack that use unknown vulnerabilities.

The latter approach relies on statistical methods to differentiate normal and anomalous behaviors. Thus, deviation from the normal behavior corresponds to anomalous one. The accuracy of this approach can be lower than the signature-based approach, however, the main advantage of this approach is a robustness against new types of anomalies. Also, this approach is suitable for high-speed networks because it is robust against packet sampling.

In this talk, we first review past literature in anomaly detection in Internet as described above. Then, we introduce our past and recent activities to tackle this problem. As a basic principle, we do not believe that there is one perfect anomaly detection algorithm for Internet backbone traffic. Therefore, we have been developing several anomaly detectors based on different theoretical backgrounds. We will briefly explain these algorithms (e.g., Multi-scale histogram, PCA, and image processing approach).

However, it is a natural idea to combine several anomaly detection algorithms to increase the accuracy of the anomaly detection, assuming that there is no one perfect algorithm. Based on this principle, we developed a new anomaly detection framework for Internet backbone traffic. The key idea of this is to construct a graph representation of anomalous events; the node is an anomalous event detected by each algorithm, and the link weight between them is the number of packets belonging to them. We apply a community mining algorithm to extract dense parts in the graph, meaning that anomalous behaviors are highlighted as well-connected sub graphs.

We demonstrate the performance improvement of this approach by using publicly available MAWI traffic traces. This longitudinal dataset since 2000 contains a wide variety of network events, so they are suitable to show the effectiveness of our approach. Indeed, we have been updating our anomaly database that is an annotation to the WIDE traffic traces by our detection algorithms for other researchers (http://www.fukuda-lab.org/mawilab). Thus, if other researchers develop a new anomaly detection algorithm, they can easily compare the performance of their algorithm to our benchmark results.

Finally, we introduce our recent activity on a new type of Internet sensor, DNS backscatter, for anomaly detection. One problem of Internet traffic anomaly detection is a locality of measurement point. Network operators know their network status by observing their networks. However, it is not easy to understand network-wide events (e.g., scans and spams). In this end, we have been developing a new Internet sensor based on DNS. The basic idea of this sensor relies on the fact that a source of a network-wide event triggers a large number of reverse DNS queries at target, though each queries have a small amount of information. Thus, we could detect anomalous events from the DNS queries with the help of collective knowledge. We show the effectiveness of DNS backscatter with DNS logs measured at Root and JP DNS servers.

- 708 -
Approaching Rate Distortion Bound with Reinforcement Message Passing

Takumi Fujita, Koji Okino and Tatsuto Murayama

Abstract—This work reviews the methods for lossy coding of Bernoulli(1/2) source with message passing algorithms/heuristics being imposed some reinforcement operations for an optimal convergence. In particular, we consider a kind of feedback technique and presents the optimal interval for the accurate decoding of the original binary sequence. By increasing the intensity of the reinforcement control, we always observe a drastic breakdown of the system performance at a certain level of feedback.

1. Introduction

Since the start point of information theory by Shannon, lossy encoding of binary information gathers attention to the wide range of mathematicians, engineers, and practitioners. However, at least when it comes to the practical point of view, it had been quite difficult to make a good pair of encoder and decoder for the Bernoulli(1/2) source, i.e., a class of the purely random sequences, until some breakthrough was made with what we call the message passing methods today [1, 2, 3]. In these techniques, we need to impose some reinforcement conditions for the equation system to converge. In this paper, we focus on the role of feedback operations in one of the earliest methods among such algorithms/heuristics to think over the physics behind the convergence.

2. System model

Let \( J \) be an \( M \)-bit source sequence, \( \xi \) an \( N \)-bit codeword, and \( \hat{J} \) an \( M \)-bit reproduction sequence, respectively. Being given a distortion \( D \) and a randomly-constructed Boolean matrix \( A \) of dimensionality \( M \times N \), we look for the \( N \)-bit codeword sequence \( \xi \), which satisfies \( J = A\xi \mod 2 \), where the fidelity criterion \( D = E[d(J, \hat{J})] \) holds. We suppose that binary alphabets are drawn randomly from a non-biased source and that the Hamming distortion measure is selected for the fidelity criterion, where the matrix is characterized by \( K \) ones per row and \( C \) per column, i.e., \( K \) and \( C \) define a particular generator matrix.

3. Algorithm/Heuristics

We proposed a feedback systems model for generating the proper codewords without what we call the dimension curse [1]. The newly introduced variables \( m_{i\mu}(t), \hat{m}_{i\mu}(t) \in [-1, +1] \) emulate the density evolution with

\[
\hat{m}_{i\mu}(t + 1) = \tanh(\beta J_i) \prod_{\nu \in \mathcal{L}(i)} m_{i\nu}(t) \tag{1}
\]

and

\[
m_{i\mu}(t) = \tanh\left( \sum_{\nu \in \mathcal{M}(i)} \tanh^{-1} \hat{m}_{i\nu}(t) + \tanh^{-1} \gamma m_i(t) \right). \tag{2}
\]

The above pair of equations (1) and (2) give an iterative procedure for codeword generation with

\[
m_{i}(t) = \tanh\left( \sum_{\mu \in \mathcal{M}(i)} \tanh^{-1} \hat{m}_{i\mu}(t) + \tanh^{-1} \gamma m_i(t) \right). \tag{3}
\]

Here, we write the set of codeword indexes \( i \) that participate in the source index \( \mu \) by \( \mathcal{L}(\mu) = \{ i \mid a_{i\mu} = 1 \} \) with \( A = (a_{i\mu}) \). Similarly, we also denote another set \( \mathcal{M}(i) \) such that it defines the set of source indexes linked to the codeword index \( i \). Practical encoding procedure for this compression model would be as follows. First, given the source sequence \( J \), we just translate the Boolean alphabets \([0, 1]\) into that of Ising ones \(\{+1, -1\}\). Then, for a certain good pair of control parameters, \( \beta \) and \( \gamma \), the equations (1) and (2) with (3) are recursively calculated until they converge to a fix point. Finally, according to the equation (3), we calculate the codeword sequence \( \xi \) by the Boolean translation.

4. Feedback Induced Order

Numerical experiments show that the algorithm with optimal parameter selection can achieve the bound for sparse construction of the codes [1]. Here, the optimal selection implies a good pair of \( \beta \) and \( \gamma \) with respect to the system performance measure, i.e., the resulting distortion \( D \). It has been already reported and widely known that the optimal value of \( \beta \) should be determined by a set of saddle
Figure 1: Empirical Performance: The feedback parameter $\gamma = 0.008$ gives the smallest distortion $D$ around $0.08$ for $K = 2$ and $C = 3$. Numerically we observe feedback induced order within the interval of $[0.008, 0.085]$. However, beyond the upper bound $\gamma = 0.085$, the system starts to converge to a non-optimal state.

point equations in the corresponding replica analysis [4]. However, when it comes to the best value of $\gamma$, we still don’t have any theoretical background for the parameter selection problem. Therefore, in this paper, we examine the systematic response induced by the existence of such feedback $\gamma$. More precisely, by imposing the optimal value $\beta = 2.35$ for the case, we numerically measures the distortion $D$ for an interval of $\gamma$. Figure 1 shows our results for the investigation. Notice here that we find an interval for nearly achieving the Shannon limit, beyond which the system suddenly loses control of decoding the original information.

5. Conclusion

In this paper, we focus on the system level effect induced by the feedback term. According to our preliminary study, it is likely that we typically observe an abrupt change of system performance at a critical value for the feedback. Future research includes the verification of this assumption using much larger scale simulations.

Acknowledgments

The authors would like to thank members in the laboratory for information theory and coding for their fruitful suggestions and comments.

References


Network Coding and Dynamical Systems

Yusuke Iseki, Koji Okino and Tatsuto Murayama

†Project Management Promotion Department, Project Management Division, Hitachi Solutions, Ltd.
‡Information Technology Center, University of Toyama
§Graduate School of Science and Engineering, University of Toyama

Email: yusuke.iseki.kf@hitachi-solutions.com, okino@itc.u-toyama.ac.jp, murayama@eng.u-toyama.ac.jp

Abstract—This work considers a kind of empirical tradeoff in network coding for dynamical systems, offering some insights for real world applications. In particular, we optimize its computational cost and the transaction delay in the network that is caused by the use of our recursive method. Future application would be also mentioned from a point of view of information network sciences.

1. Introduction

The rise of the notion “Internet of Things” has been gathering attention to the wide range of network engineers and computer scientists. At the same time, it is becoming crucial to consider how to network many devices with the limited computational resources. One of the conventional approaches would be the naive application of what is called network coding in the area of information theory. This classical method provides an efficient procedure to broadcast digital contents to many subscribers at the largest data rate possible. However, it is also known that practical encoding for typical large networks is not that easy. In this paper, we propose a recursive construction of network coding based on a graph partitioning scheme, and try to numerically optimize their sizes to enhance the overall data rate under the same computational resources.

2. System model

Network coding is a kind of “store and forward” type routing technology which enables us to deal with multiple inputs and outputs efficiently. The basic concept was first proposed by Ahlswede, Cai, Li, and Yeung at the outset of the 21st century [1]. After that, Li, Yeung, and Cai offered the method of linear network coding as a practical framework to realize the new routing technology [2]. Together with the linear information flow (LIF) algorithm given by Jaggi, Sanders, Chou, Elfros, Egner, Jain, and Tolhuizen, it is possible for us to implement the basic encoding of network coding into real world networks [3]. In this paper, we numerically examine a class of ideal networks with the following assumptions.

- We consider the 36-layered butterfly-type network as a system model.
- We use the LIF-algorithm to encode all network transactions.
- We apply simple “wait and throw” principle for the packet collision.

Notice that we choose the butterfly network as a building block for our network decomposition approach, since using this element enables us to find the trivial solution to the packet collision problem in the layered model. The details of the decomposition will be reported elsewhere.

3. Numerical Results

Figure 1 shows the empirical performances of the overall coding time and the system level delays given a depth of the layered network. As is expected, an increase of the element network size increases the overall coding time and decreases the systematic delay steps.

Figure 2 represents the empirical tradeoff between the two measures. In general, we prefer shorter coding time \(x_t\) and fewer delay steps \(x_d\). Write

\[ \beta(x_t, x_d) = x_t + \alpha x_d \quad (0 < \alpha < \infty). \]
Figure 2: Empirical Tradeoff: Solid line shows Numerical tradeoff between coding time and delay steps. The transaction delay decreases as the coding time increases meaning that we use larger butterfly units for the network decomposition.

Here the parameter $\alpha$ denotes our preference for the two measures. By choosing $\alpha = 0$ we have $\beta(x_t, x_d) = x_t$, which corresponds to the user who prefers shorter coding time needing to reconfigure the dynamical network every minute. On the other hand, if we consider $\alpha \to \infty$ then $\beta(x_t, x_d)$ equals $\alpha x_d$ and this implies the user who does not like any delays in the transactions. That is, by imposing the value of $\alpha$ for the potential user preference, we could decide the optimal depth level to enhance the system performance.

4. Conclusion

Since there would be so many mobile devices connecting to the internet already, we need to regard networks as dynamical systems rather than static ones. In this paper, we reconsider computational cost of its overhead processes required to realize the typical network coding from a point of view of the standard pareto optimization. Natural directions for future research include generalizing the network architecture and the theoretical sophistication.

Acknowledgments

The authors would like to thank members in the laboratory for information theory and coding for their fruitful suggestions and comments.

References


Abstract—The shortest path problem is one of the fundamental optimization problems, which has been well investigated extensively so far from not only theoretical but also practical viewpoints. The shortest path problem can be solved by the Dijkstra algorithm, which is very efficient algorithm; however, even the Dijkstra algorithm is not sufficient for recent applications that require the problem to be solved in a very short time, if the algorithm is applied to a very large network with a huge number of vertices and edges such as the road networks. Therefore, we need the algorithms to find more quickly shortest paths many time. An approach is the combination of preprocessing and shortest path queries, and some algorithms based on this approach have been proposed so far. In this paper, we propose a method to find an approximation path. An approximation solution is sufficiently useful for practical use such as the route search in a road network. We focus on the geographically inhomogeneous density of vertices. The basic idea of the proposed method is the clustering based on the inhomogeneous density and the reduction of the network size by shrinking the clusters. The shrunk network is constructed by preprocessing; when a query occurs, the shortest path is searched in the shrunk network. Thus, we can find an approximation path much faster.

In this paper, by using the approach of the combination of preprocessing and queries, we propose a method to find an approximation path. We focus on the geographically inhomogeneous density of vertices, and the basic idea of the proposed method is the clustering based on the inhomogeneous density and the reduction of the network size by shrinking the clusters. The shrunk network is constructed by preprocessing; when a query occurs, the shortest path is found in the shrunk network. Thus, we can find an approximation path much faster, although it is not always the shortest path much faster.

1. Introduction

The shortest path problem is one of the fundamental optimization problems, which has been well investigated extensively so far from not only theoretical but also practical viewpoints. For example, the on-line route search on a map is widely and frequently used. This means that an algorithm to solve the shortest path problem is widely and frequently executed in our daily life.

The shortest path problem can be solved by the Dijkstra algorithm [1], which is very efficient algorithm; however, even the Dijkstra algorithm is not sufficient for recent applications that require the problem to be solved in a very short time, if the algorithm is applied to a very large network with a huge number of vertices and edges such as the road networks. Therefore, we need the algorithms to find more quickly shortest paths many time.

An approach is the combination of preprocessing and shortest path queries. As the networks such as road networks does not change frequently, we can design the data structure by preprocessing to answer the query quickly. It can be expected to reduce the computational time drastically when a query occurs. Some algorithms based on this approach such as the hierarchical mesh sparse method have been proposed so far. We can find the shortest path by using these algorithms. In an actual application, even an approximation solution is sufficiently useful for practical use such as the route search in a road network. It can be expected to find an “approximation path” that is not always the shortest path much faster.

2. Related Works

The methods based on the combination of preprocessing and shortest path queries have been investigated; for example, transit node routing [2], highway hierarchy [3], and level-wise mesh sparsification [4].

The transit node routing method [2], in the preprocessing phase, chooses the transit vertices such that, for any pair of vertices, at least a transit vertex is included in the shortest path, and finds the shortest paths of all pairs of the transit vertices. When a query occurs, from a transit vertex near the source to a transit vertex near the destination.

The highway hierarchy method [3] focuses on the edges frequently used. These edges are used as a highway for a path to a farther vertex, and a path between two vertices is searched in the usually sparse highway network.

The level-wise mesh sparsification method [4] is similar to the highway hierarchy. In the preprocessing phase, it makes the level-wise sparse networks based on the geometric partition. When a query occurs, the shortest path is searched in the level-wise sparse networks.

Shortest-path queries by using property of geographical density in road network
Yukitoshi Senshu† and Hiroyoshi Miwa†

†Graduate School of Science and Technology, Kwansei Gakuin University
2-1 Gakuen, Sanda, Hyogo 669-1337, Japan
Email: senshu@kwansei.ac.jp, miwa@kwansei.ac.jp
3. Shortest-path queries by using property of hierarchy structure of geographical density

The graph representing a network is $G = (V, E)$. Here, $V$ and $E$ are the set of vertices and the set of edges of $G$, respectively, the number of vertices is $|V| = n$, and the number of edges is $|E| = m$. We associate the weight with each edge by weight function $w : E \rightarrow \mathbb{R}^+$. The path length is defined as the sum of the weight of all the edges included in the path. The shortest path between two vertices is defined as the path with the smallest path length among all paths between the vertices. When we associate the weight with each vertex by extended weight function $w : V \cup E \rightarrow \mathbb{R}^+$, the path length is defined as the sum of the weight of all the edges and the vertices included in the path.

We assume that a network $N = (G = (V, E), w)$ is embedded on a 2-dimensional plane and each vertex has 2-dimensional coordinates. A geometric network such as a road network and a railroad network has a 2-dimensional coordinates. Let $x(v)$ and $y(v)$ be the $x$- and $y$-coordinates of $v \in V$, respectively, and let $p$ be the function $p : V \rightarrow \{(x(v), y(v)) \in \mathbb{R}^2\}$.

The shortest path query problem is the problem to output the shortest path and its length, when a shortest path query $(s, t)$ where $s, t \in V$ occurs. A naive approach for this problem is a method that always executes the Dijkstra algorithm for the whole network for every query. However, it needs too much computational resource and time in case of a large network. We describe the preprocessing approach. When a network $N = (G = (V, E), w, p)$ is given, the network is transformed by the preprocessing phase, and we get and store the transformed network $N' = (G' = (V', E'), w')$. When a shortest path query $(s, t)$ where $s, t \in V$ occurs, an algorithm searches and output the shortest path and its length in the network by using $N'$. $N'$ is generally a sparsified network whose number of vertices and edges are much smaller than the original network; therefore, we can expect the computational time of the preprocessing approach is much smaller than the naive approach.

We propose two algorithms in this paper as follows. We describe the basic idea of the algorithms. The algorithm divides a given network into the clusters based on the geographical location by using the single linkage method [5]. As there are many roads inside an urban area, the cross over points are close to each other in narrow area, that is, the density of the cross over points is high; on the other hand, the density of the cross over points is low in the suburb and the area connecting cities. The cluster division by the single linkage method extracts the high density areas. The network is constructed by shrinking each cluster into a vertex. The size of the shrunk network is much smaller than the original network. We can expect that the shortest path in the shrunk network is “close” to the shortest path in the original network; especially, when the source and destination vertices are distant each other, because the geographical direct distance is dominant in the lengths of the paths between two vertices. The shrunk network is constructed in the preprocessing phase; when a query occurs, the shortest path is searched in the shrunk network, then the shortest path in the original network is reconstructed from the path.

The proposed algorithm **SP-Average** uses the average of the path lengths for all pairs of vertices in a cluster as the weight of a vertex in the shrunk network. We show an example of this algorithm in Figure 1.

![Figure 1: Concept of algorithm SP-Average.](image)

We show the algorithms the preprocessing phase in Algorithm 1 and the query phase in Algorithm 2 as follows.

**Algorithm 1: Algorithm SP-Average-Preprocessing**

**Input:** A network $N = (G = (V, E), w, p)$.

**Output:** The sparsified network $N' = (G' = (V', E'), w')$ where $w' : V' \cup E' \rightarrow \mathbb{R}^+$.

1. Make the clusters $\{V_1, V_2, \ldots, V_k\}$ by using the single linkage method.
2. for $i = 1$ to $k$
3. | Find and store the shortest paths and the path lengths for all pairs of vertices in $V_i$.
4. | Calculate the average $a_i$ of the path lengths for all pairs of vertices in $V_i$.
5. Construct the graph $G' = (V', E')$ that $V' = \{v_1, v_2, \ldots, v_k\}$ and $E' = \{(v_i, v_j) \mid \exists (u, v) \in E, u \in V_i, v \in V_j\}$.
6. for $e \in E'$ do
7. | Let $w'(e) = w((u, v))$ where $u \in V_i, v \in V_j$.
8. for $i = 1$ to $k$
9. | Let $w'(v_i)$ be $a_i$.
10. return the network $G' = (V', E'), w')$.

The other proposed algorithm **SP-Clique** substitutes a
Algorithm 2: Algorithm SP-Average-Search

Input: A query \((s, t)\) where \(s, t \in V\) in \(N = (G = (V, E), w, p)\).

Output: The shortest path and its path length in \(G\).

1. Find \(V_i\) including \(s\) and \(V_j\) including \(t\).
2. Find the shortest path \(P\) in the network \((G' = (V', E'), w')\) between \(v_i\) and \(v_j\).
3. Find the shortest path in each \(V_k\) of \(G\) where vertex \(v_h\) of \(G'\) is included in \(P\), and connects all the paths.

return the shortest path and its length.

4. Performance Evaluation

We evaluate the performance of the algorithms proposed in section 3.

We generate a grid graph, and the coordinates of the vertices are determined so that they are located at the grid point first and then uniformly randomly distributed around. We locate some networks (we refer this as sub-networks hereafter) apart and by connecting the networks by edges. Thus, we generate a network with the number of vertices of 10000 including sub-networks corresponds to clusters.

We compare the performance of the case that the Dijkstra algorithm is applied to an original network for all queries (Dijk), the case SPA that SP-Average-preprocessing and the case SPC that SP-Clique-preprocessing for queries are applied, and the case SPA that SP-Clique-preprocessing and SP-Clique-Search for queries are applied. The number of queries is 1000, and the source and destination vertices are randomly chosen. We used Dell Latitude E5500 (Intel Core i5-4130, 2.00GHz and 2.59GHz, 4GByte RAM) for the numerical experiments.

First, we show the computational time for a network whose sub-networks is 16 in Table 1. We observe how the proposed algorithms depends on the number of clusters.

The results in Table 1 shows that SPA is much faster than Dijk. SPC is faster than Dijk, when the clusters are appropriately divided.

Both SPC and SPA do not always output the shortest
Table 1: Computational time (ms) and number of sub-networks.

<table>
<thead>
<tr>
<th>Num. sub-nets</th>
<th>16</th>
<th>8</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dijk</td>
<td>133360</td>
<td>133360</td>
<td>133360</td>
<td>133360</td>
</tr>
<tr>
<td>SPC</td>
<td>13468</td>
<td>27500</td>
<td>63796</td>
<td>89001</td>
</tr>
<tr>
<td>SPA</td>
<td>5625</td>
<td>15969</td>
<td>39218</td>
<td>35813</td>
</tr>
</tbody>
</table>

We observe the number of different paths by SPC from the shortest path in Table 2.

Table 2: Number of different paths by SPC from the shortest path.

<table>
<thead>
<tr>
<th>Num. sub-nets</th>
<th>16</th>
<th>8</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPC</td>
<td>0</td>
<td>7</td>
<td>73</td>
<td>91</td>
</tr>
</tbody>
</table>

The result shows that the shortest paths are determined by SPC when the clusters are appropriately divided, but that the number of the different paths from the shortest paths increases according to the increase of the difference from the appropriate clustering.

We observe the ratio of the average of the path lengths by SPA to the shortest path length in Table 3.

Table 3: Ratio of the average of the path lengths by SPA to the shortest path length.

<table>
<thead>
<tr>
<th>Num. sub-nets</th>
<th>16</th>
<th>8</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPA</td>
<td>1.08</td>
<td>1.12</td>
<td>1.14</td>
<td>1.13</td>
</tr>
</tbody>
</table>

The ratio becomes large according to the increase of the difference from the appropriate clustering. However, the ratio does not exceed two.

The results in Table 2 and Table 3 show that the proposed algorithms can find an approximation path. Furthermore, when the clusters are appropriately divided, the approximation ratio or the number of different paths from the shortest path is small. This implies that the clustering plays an important role.

Next, we observe the results of other networks whose coordinates are different in Table 4, Table 5, and Table 6.

Table 4: Computational time (ms) and number of sub-networks.

<table>
<thead>
<tr>
<th>Num. sub-nets</th>
<th>16</th>
<th>8</th>
<th>4</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dijk</td>
<td>2500</td>
<td>4375</td>
<td>2500</td>
<td>5000</td>
</tr>
<tr>
<td>SPC</td>
<td>133124</td>
<td>133124</td>
<td>133781</td>
<td>133781</td>
</tr>
<tr>
<td>SPA</td>
<td>15969</td>
<td>9218</td>
<td>15350</td>
<td>34844</td>
</tr>
</tbody>
</table>

We proposed the method to find a path approximating the shortest path. We focus on the geographically inhomogeneous density of vertices, and the basic idea of the proposed method is the clustering based on the inhomogeneous density and the reduction of the network size by shrinking the clusters. The shrunk network is constructed by preprocessing; when a query occurs, the shortest path is found in the contracted network. Thus, we can find an approximation path much faster, although it is not always the shortest path. We proposed two algorithms based on the approach and evaluated by the numerical experiments. Both algorithms can find the paths faster than the method by searching the shortest path in the original network.

5. Conclusion

We proposed the method to find a path approximating the shortest path. We focus on the geographically inhomogeneous density of vertices, and the basic idea of the proposed method is the clustering based on the inhomogeneous density and the reduction of the network size by shrinking the clusters. The shrunk network is constructed by preprocessing; when a query occurs, the shortest path is found in the contracted network. Thus, we can find an approximation path much faster, although it is not always the shortest path. We proposed two algorithms based on the approach and evaluated by the numerical experiments. Both algorithms can find the paths faster than the method by searching the shortest path in the original network.

References


