

Graph Theoretical Reviews on Constrained Coding for Data Storage Media

Akiko MANADA

Department of Electrical, Electronics and Information Engineering
Nagaoka University of Technology

2022 年電子情報通信学会総合大会 チュートリアルセッション
AT-1：次世代ストレージに対する符号化法の最近の動向と今後の展開

March 16, 2022

謝辞：本発表は JSPS 科研費基盤研究 (C)19K11833 の助成を受けた。
発表に先立ち、貴重なコメントをいただいた専修大学の太田隆博先生に感謝いたします。

本発表の流れ

1 Introduction

- 誤り訂正符号と制約符号の違い
- 記録媒体と制約符号の関連性
- グラフ理論と制約符号の関連性

2 Constrained System とは

- Language と Graph Theory
- Constrained System と Presentation
- Shannon cover の有用性と主結果

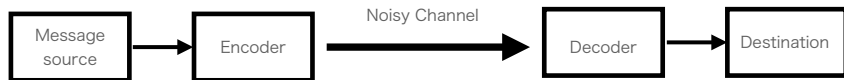
3 次世代ストレージへの応用と解決すべき問題

- DNA ストレージへの応用
- DNA ストレージに求められる制約と課題

4 Conclusion

1. Introduction

誤り訂正符号と制約符号



- 通信路での誤りがランダムに発生
→ 誤り訂正符号を用いて訂正.
- 「通信路での誤りに影響を受けやすいデータ列」が予めわかっている
→ 制約符号を用いて, そのようなデータ列を排除.

記録媒体と制約符号の関連

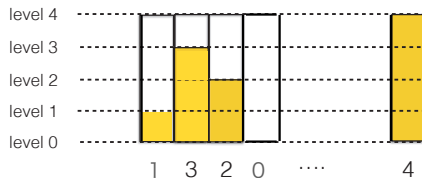
フラッシュメモリ：高密度・不揮発な記録媒体（例：USB, SSD）

メッセージの書き込み方法：

- ① 保存したいメッセージをあるデータ列 $d = d_1 d_2 \dots d_N$ に変換.
- ② i 番目のセル ($1 \leq i \leq N$) に、電荷レベルが d_i になるまで電荷を注入する.

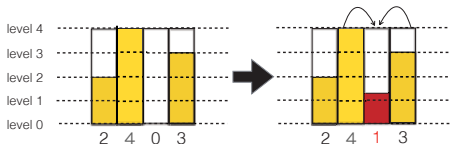
メッセージの読み出し方法：

- ① 各セルの電荷レベルを量子化することで、データ列を復元.
- ② データ列に対応したメッセージを返す.



フラッシュメモリのセル間干渉

セル間干渉：隣接するセルの電荷から干渉を受けてセルの電荷レベルが所望以上の値になり、データ列復元の際に正しくない値を返す誤り。



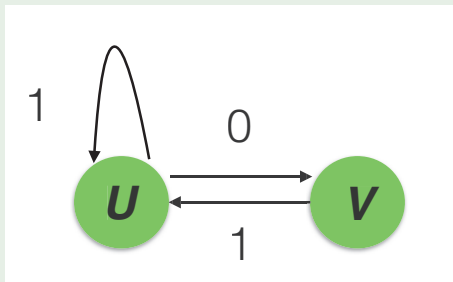
- デバイスの小型化に伴い、セル間干渉を防ぐことが重要視される。
- 隣接するセルとの電荷に差があるほど、セル間干渉を起こしやすい（セル間干渉の影響を受けやすいデータ列が知られている）。

⇒ セル間干渉に影響を受けやすいデータ列を排除する制約符号（書き込み方法）を用いて、セル間干渉に耐性を持たせるように符号化できないだろうか？

制約を満たす系列の列挙方法

Question: 00 を含まないバイナリ系列 (0 – 1 系列) を列挙せよ.

Example 1 (ラベル付き有向グラフを用いた列挙)



- 任意の経路に沿ってラベルを読むことでできるバイナリ系列は、00 を含まない.
- 00 を含まないバイナリ系列は、ある経路に沿ってラベルを読むことで生成できる.

ラベル付き有向グラフの利点

- コンパクトに全ての情報を含めることができる.
- 制約を満たす系列にエンコードする符号器を作成する際に用いられる.
- 最大符号化率を正確に、かつ簡単に求めることができる.

Constrained system と密接な繋がりがある.

2. Constrained System とは

参考文献：

- [1] D. Lind and B. Marcus, "An Introduction to Symbolic Dynamics and Coding," Cambridge University Press, 1995.
- [2] B. H. Marcus, R. M. Roth, and P. H. Siegel, "An introduction to coding for constrained systems." Unpublished Lecture Note available at <https://ronny.cswp.cs.technion.ac.il/wp-content/uploads/sites/54/2016/05/chapters1-9.pdf>
(最終閲覧日：2022 年 3 月 16 日)

Language:

- アルファベット (有限シンボル集合) : Σ
- 系列 : $\mathbf{w} = w_1 w_2 \dots w_\ell \in \Sigma^\ell$ ($\ell \geq 0$)
- 系列 $\mathbf{w}$ の部分系列 $\mathbf{w}' = w'_1 w'_2 \dots w'_{\ell'} \quad (\ell' \leq \ell)$:
ある整数 i ($1 \leq i \leq \ell - \ell' + 1$) に対して, $\mathbf{w}' = w_i w_{i+1} \dots w_{i+\ell'-1}$.

(例) 01 は 10011 の部分系列である.
「10011 は 01 を含む」ともいう.

Graph Theory:

- ラベル付き有向グラフ $G = (V, E, L)$:
 - 頂点集合 V ;
 - (有向) 辺集合 $E = \{e = (u, v) : u, v \in V\} \subset V \times V$;
 - 辺 e は頂点 u の出辺, かつ頂点 v の入辺である.
 - 辺のラベリング $L : E \rightarrow \Sigma$
- 決定的なラベル付き有向グラフ: どの頂点の出辺もラベルが異なる.

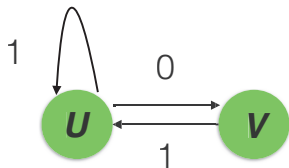
Example 2 (決定的なグラフ・決定的でないグラフ)



Graph Theory:

- 頂点集合 $V = \{v_1, v_2, \dots, v_{|V|}\}$ の隣接行列 $A = A_G$:
 - $|V| \times |V|$ の正方行列;
 - (i, j) 成分は頂点 v_i から頂点 v_j への有向辺の数

Example 3 (グラフと隣接行列)



$u = v_1, v = v_2$ と考えると,
隣接行列は $\begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}$

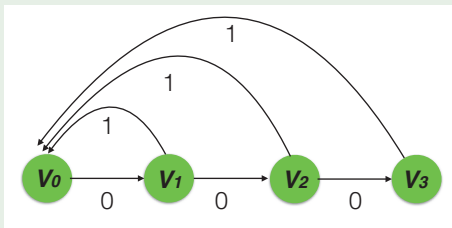
Constrained System と Presentation

Definition 1 (Constrained System $\mathcal{S} = \mathcal{S}_G$)

Constrained system $\mathcal{S} = \mathcal{S}_G$ とは、あるラベル付き有向グラフ G の経路に沿ってラベルを読んでいくことで生成される系列の集合である。このとき、 G を $\mathcal{S}$ の **presentation** と呼ぶ。

Example 4 ((1,3)-Run Length Limited (RLL) constrained system)

「1 と 1 の間の 0 の数は 1 以上 3 以下」かつ「0 の連長は最大 3」とする。



Constrained System と 禁止語との関連

Remark 1 (Constrained system の禁止語表現)

- Constrained system $\mathcal{S} = \mathcal{S}_{\mathcal{F}}$ は, 禁止語集合 $\mathcal{F} = \{f_1, f_2, \dots\}$ 内の系列を含まない系列の集合として考えられる [Chapter 3.1, [1]]
- 禁止語集合 $\mathcal{F}$ が有限ならば, $\mathcal{F}$ 内の系列を含まない系列の集合 $\mathcal{S}_{\mathcal{F}}$ は Constrained system である [例: [3, 4]]

Example 5 ((1,3)-Run Length Limited (RLL) constrained system)

「1 と 1 の間の 0 の数は 1 以上 3 以下」かつ「0 の連長は最大 3」とする制約なので,

$$\mathcal{F} = \{11, 0000\}$$

としたバイナリ系列全体と考えられる.

[1] D. Lind and B. Marcus, "An Introduction to Symbolic Dynamics and Coding," Cambridge University Press, 1995.

[3] N.-C. de Bruijn, "A combinatorial problem," Koninklijke Nederlandse Akademie V. Wetenschappen, vol. 49, no. 7, pp. 758–764, 1946

[4] M. Crochemore, F. Mignosi, and A. Restivo, "Automata and forbidden words," Information Processing Lett., vol. 67, no. 3, pp. 111–117, 1998.

Remark 2 (決定的な presentation の存在)

Constrained system には, 必ず決定的な presentation が存在する.

Definition 2 (Shannon Cover)

Constrained system S の決定的な presentation のうち, 頂点数が最小のものを S の **Shannon cover** とよぶ.

Shannon cover の有用性

Constrained system $\mathcal{S}$ の Shannon cover は,

- $\mathcal{S}$ を表現する最も標準的なプレゼンテーションとして用いられる.
- 任意の系列を $\mathcal{S}$ の系列に符号化する符号器を作成する際に用いられる [Chapter 5, [1]].
- 任意の系列を $\mathcal{S}$ の系列に符号化する際の最大符号化率 $C(\mathcal{S})$ は,

$$C(\mathcal{S}) = \log \lambda_A$$

で求められる. ここで λ_A は, $\mathcal{S}$ の Shannon cover の隣接行列の最大固有値とする [Chapter 4, [1]].

Shannon cover の解析は constrained system の理論解析の核である.

[1] D. Lind and B. Marcus, "An Introduction to Symbolic Dynamics and Coding," Cambridge University Press, 1995.

Constrained System に関するこれまでの主結果

- Crochemore らの手法 [4] で得られた決定的 presentation が Shannon cover になるのはどのような Constrained system か？ [5].
- Constrained System が規約な Presentation を持つ条件 [6].
- Shannon cover が一意になるのはどのようなケースか？ [7].
- ゼータ関数（周期系列を表現する母関数）の求め方 [8]

[4] M. Crochemore, F. Mignosi, and A. Restivo, "Automata and forbidden words," Information Processing Lett., vol. 67, no. 3, pp. 111-117, 1998.

[5] A. Manada and N. Kashyap, "On the Shannon covers of certain irreducible constrained systems of finite type," Proc. 2006 IEEE Int. Symp. Inform.Theory, 2006, pp. 1477-1481.

[6] Tetsuya Kobayashi, Akiko Manada, Takahiro Ota and Hiroyoshi Morita, "On the Irreducibility of Certain Shifts of Finite Type," IEICE Trans. Fundamentals, Vol. E96-A, No. 12, pp. 2415-2421, Dec. 2013

[7] A. Manada, "On a Shannon cover of certain reducible shift of finite type," in Proc. 2012 International Symposium on Information Theory and Its Applications, 2012, pp. 606-610.

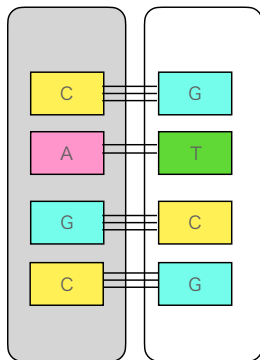
[8] A. Manada and N. Kashyap, "On the Zeta Function of a Periodic-Finite-Type Shift," IEICE Trans. Fundamentals, Vol. E96-A, No. 6, pp. 1024-1031, June 2013.

3. 次世代ストレージへの応用と 解決すべき問題

DNA ストレージへの応用 1/3

DNA (DeoxyriboNucleic Acid：デオキシリボ核酸)

- アデニン (A), チミン (T), シトシン (C), グアニン (G) の4つの塩基から構成される核酸.
- 生物に関する遺伝情報（成長，発達等）を保持・伝達する



DNA へのデータ保存：

- ① データを 4 値 (A, T, C, G) の DNA 系列に変換する
- ② 10^2 から 10^3 の長さの DNA 鎖に切断し，番地情報をつける．
- ③ 保存液に入れて保管．その際，同じ DNA 鎖のコピーも作成する．

DNA からのデータ読み出し：

- ① 取り出したい情報の番地持つ DNA 鎖を取り出す．
- ② DNA 鎖の系列を読み取る．
- ③ 読み取った系列から元のデータを復号する．

DNA ストレージへの応用 3/3

Table: DNA ストレージと他の記録媒体との比較 (文献 [9] より引用)

ストレージの種類	HDD	フラッシュメモリ	DNA
読み書きスピード (μ s per bit)	$\sim 3,000 - 5,000$	~ 100	< 100
データ保存期間 (years)	> 10	> 10	> 100
電力使用量 (watts per gigabyte)	~ 0.04	$\sim 0.01 - 0.04$	$< 10^{-10}$
データ密度 (bits per cm^3)	$\sim 10^{13}$	$\sim 10^{16}$	$\sim 10^{19}$

[9] Andy Extance, "How DNA could store all the world's data," Nature, vol. 537, no. 7618, pp. 22-25, 2016.

DNA ストレージでの制約

DNA 鎖における主要な制約（参考文献 [10, 11]）

- 連長制約：同じ塩基は連続して高々 k 個までとする.

$k = 5$ のとき,

- ○: AAAGGCAGAACGGCAT
- ×: AAAAAAATAAGGCAT

- GC-バランス制約：「G と C の総数 = A と T の総数」とする.

- ○: AAAGGCAGAACGGCAT
- ×: AAAAAAATAAGGCAT

[10] S. Yazdi, H. Kiah, E. Garcia, J. Ma, H. Zhao, and O. Milenkovic, "DNA-based Storage: Trends and Methods," IEEE Trans. Molecular, Biological, Multi-Scale Commun., vol. 1, no. 3, pp. 230-248, 2015.

[11] Y. Erlich and D. Zielinski, "DNA Fountain enables a robust and efficient storage architecture," Science, vol. 355, no. 6328, pp. 950-954, 2017.

GC-バランス制約を満たす記号列の集合は, Constrained system ではない.

つまり, GC-バランス制約を満たす系列を列挙するような presentation は存在しない.

- 理由: 禁止語が考えられない
(どんな系列も GC-バランス制約を満たす系列の部分列になりうる).

GC-バランス制約について 2/2

GC-バランス制約のみ考慮する場合には、(constrained system は用いずとも) 効率的なやり方は提示されている
(例: Knuth の方法 [12])

しかし、GC-バランス制約と他の制約を組み合わせる場合には、一筋縄ではいかない (数え上げ手法を用いるのが主流か [13, 14])

Theorem 1 ([Theorem IV.3, 14])

バイナリ系列において、連長制約とバランス制約を満たしている記号列全体の最大符号化率は、連長制約を満たす記号列全体の最大符号化率と等しい。

[12] Donald E. Knuth, "Efficient Balanced Codes," IEEE Trans. Inform. Theory, vol. IT-31, no. 1, pp. 51-53, 1986.

[13] Kees A. Schouhamer Immink and Jos H. Weber and Hendrik C. Ferreira, "Balanced Runlength Limited Codes Using Knuth's Algorithm," Proc. of 2007 IEEE International Symposium on Information Theory, pp. 1686-1690, 2021.

[14] A. Manada and H. Morita, "On the Capacities of Balanced Codes with Run-Length Constraints," Proc. of 2017 IEEE International Symposium on Information Theory, pp. 1391-1395, June 2017.

多次元拡張の難しさ

より高密度に保存したい → 2次元（多次元）でのデータ保存.

しかし、二次元配列における制約符号を考えることは、非常に困難であり、様々な問題が未解決である. 実際、与えられた禁止パターンに関して

- 禁止パターンを含まない二次元平面は存在するか？
- 禁止パターンを含まない周期的二次元平面は存在するか？

は、「有限回のアルゴリズムで解けない」と示されている (swamp of undecidability) [15].

[15] D. Lind, "Multi-Dimensional Symbolic Dynamics," Proc. of Symposia in Applied Mathematics, pp. 61-80, 2004.

有限長解析の必要性

問題点をまとめると,,,

- グラフで表現できないケースはどうするか？
- 実際のデータは有限長で扱うため、無限長に固執する必要はそこまではないか？

有限長での

- 禁止パターンを含まない配列列挙（数え上げ）
- 符号化方法の提示とその有用性

等についてコツコツと積極的に行うことも必須。ひいては、有限長の結果が無限長解析の突破口となるのでは？

DNA ストレージに起こる誤り

- 突然変異（置換誤り）：塩基が他の塩基に変化する.
- 遺伝子の交叉（挿入・削除誤り）：DNA 鎖が途中で切れたり，切れた部分に他の DNA 鎖が挿入される.

誤りに耐性を持たせる必要があるが，

- 制約符号は「誤りを起こりにくくする」が「100% 誤りが起こらない」わけではない.
- 制約符号は誤り検知・訂正能力はない（誤りを増幅させる可能性がある）

誤り訂正符号との相性 2/2

誤り訂正符号（削除・挿入誤り）との融合は必須.

- ラン長制約と誤り訂正符号（例：[16]）
- ラン長制約，誤り訂正符号と GC-バランス制約（例：[17]）

Constrained system with unconstrained positions [18]（どのシンボルを入れても制約を満たす系列となる場所）からのアプローチ？

Example 6 (00 を禁止語とする場合の unconstrained positions)

1□1□1□...1□

[16] R. Takemoto and T. Nozaki, "Construction and Encoding Algorithm for Maximum Run-Length Limited Single Insertion/Deletion Correcting Code," IEICE Trans. Fundamentals of Electronics, Communications and Computer Sciences, E105-A, no. 1, pp. 35-43, 2022.

[17] T. T. Nguyen, K. Cai, K. A. S. Immink, and H. M. Kiah, "Capacity-approaching constrained codes with error correction for DNA-based data storage," IEEE Trans. Inform. Theory, vol. 67, no. 8, pp. 5602-5613, 2021.

[18] J. Campello de Souza et al., "Constrained systems with unconstrained positions," IEEE Trans. Inform. Theory, vol. 48, no. 4, pp. 866-879, 2002.

4. Conclusion

本発表のまとめ

本発表の概要

- 制約符号と記録媒体との関連性
- Constrained system とグラフ理論の関係性
- DNA ストレージに対する制約符号の問題点・課題

今後の研究方針

- 有限長での理論的解析
- シミュレーションを用いた実験と予測

グラフとこれらの結果を融合させ、記録媒体の研究へ貢献していきたい。

THANK YOU VERY MUCH FOR YOUR ATTENTION!