ランダムなしりとりの平均場解析

藤田悠朔 [1], 鈴木岳人 [2], 水口毅 [1]

[1] 大阪公立大学 大学院理学研究科 物理学専攻
[2] 高千穂大学 人間科学部

概要

しりとりは前の単語に、その末尾文字から始まる単語をつなげるゲームである。本研究ではしりとりで使用できる単語を定め、その中からランダムに単語を選ぶ過程を考える。この過程はネットワーク上の自己回避ランダムウォークとみなすことができる。我々は単語列の長さ（鎖長）に着目し、平均場近似を施した場合の鎖長分布に関する理論的な解析を行った。

Mean-field Analysis of Random Word Chain Games

Yusaku Fujita [1], Takehito Suzuki [2], Tsuyoshi Mizuguchi [1]

[1] Department of Physics, Graduate School of Science, Osaka Metropolitan University
[2] Faculty of Human Sciences, Takachiho University

Abstract

A word chain game is a game where players connect words beginning with the last character of the previous word. In this study, a “dictionary”, namely a set of usable words in the game, is defined, and a random word-selecting process in the dictionary is considered. This process can be regarded as a self-avoiding walk on the dictionary network. We focused on the chain length of each trial and analysed its distribution with a mean-field approximation.

1　はじめに

しりとりは、“りんご”→“ごりら”→“らっぱ”→ ··· のように、前の単語にその末尾文字から始まる単語をつなげていくゲームである。他のルールとしては、(i) 一度使用した単語は再度使用できない (ii) 次の単語がなくなったら終了（負け）がある。しりとりの先行研究として、使用できる単語を決めたときに最長で何単語続くのか [1] や、2 人しりとりにおける最も効果的な戦略は何か [2] などが知られている。いずれも戦略的にしりとりの鎖長（単語列の長さ）を変化させる研究である。我々は戦略を持たないしりとりの鎖長に着目した。

本研究では、使用できる単語の集合（辞書）を定め、ゲームの戦略を度外視したランダムなしりとりを考える。すなわち、単語は残されている選択可能な単語の中からランダムに選び、選択された単語は辞書から消す、という過程を終了するまで繰り返す。こうしてできる鎖長の分布を数値的および理論的に解析した。

図 1 は文学作品 “Moby-Dick” [3] に登場する英語の名詞 19088 語で 5 万回ランダムなしりとりを実行したときの鎖長分布である。非対称で歪な形状の分布が得られた。また、5 万回の試行において終了文字（最後の単語の末尾文字）として現れたのは x と y のみであった。図 2 は終了文字ごとの鎖長分布であり、文字ごとに分布の形状が異なっている。それぞれの分布はどのようにして得られるのかという疑問が提起される。

Moby-Dick 辞書には 26 文字が使われており、分布を求めるのは容易ではない。本論文ではランダムなしりとりをネットワーク上の自己回避ランダムウォー

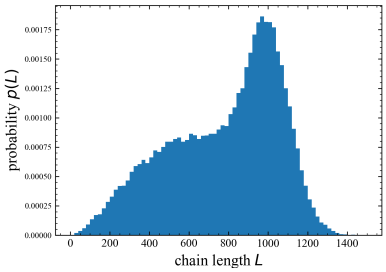


図 1: Moby-Dick 辞書の鎖長分布（50000 回試行）

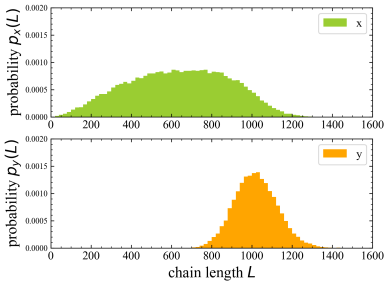


図 2: Moby-Dick 辞書の終了文字ごとの鎖長分布（上段：x, 下段：y, 50000 回試行）

クと捉え、ネットワーク構造から導かれる終了文字の条件を紹介する。さらに、使用されている文字の数が 2 という単純な場合に限定し、その鎖長分布について、平均場近似を用いた解析結果を報告する。

2　辞書ネットワーク

辞書の単語数を D、しりとりに使われる文字の種類数を C とする。すべての文字を頂点にとり、任意の文字 θ, ϕ について、θ で始まり ϕ で終わる単語が n 個あれば、頂点 θ から頂点 ϕ への有向辺を n 本張る。こうして辞書から頂点数 C、リンク数 D の多重有向グラフを構成することができる。図 3 は Moby-Dick 辞書の辞書ネットワークである。

ランダムなしりとりは辞書ネットワーク上の通過した辺を回避する自己回避ランダムウォーク (SAW) に対応する。これまで正方格子上 [4] や単純ネットワーク上 [5] [6] で SAW を行ったときの経路長分布が数値的または理論的に求められている。ただし、これらの研究で扱っている SAW は辺ではなく頂点を回避する SAW であることに注意が必要である。

終了文字となるための条件を考えたい。文字 θ の入次数を $k_{\mathrm{in},\theta}$、出次数を $k_{\mathrm{out},\theta}$ とする。θ が図 4 (a) $k_{\mathrm{in},\theta} < k_{\mathrm{out},\theta}$ を満たす場合、θ に入って出ていくを繰り返した結果、入次数が先に 0 となり θ で終了す

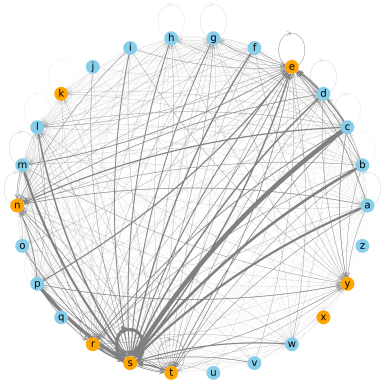
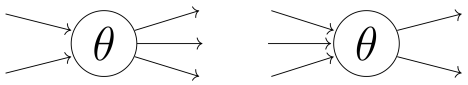


図 3: Moby-Dick 辞書の辞書ネットワーク。辺の太さは有向辺の数に比例しており、頂点の色は橙が終了可能文字を、青が終了不能文字を表している。

ることはない。一方で、図 4 (b) $k_{\mathrm{in},\theta} > k_{\mathrm{out},\theta}$ を満たす場合、出次数が先に 0 となり θ で終了する可能性が生じる。また、$k_{\mathrm{in},\theta} = k_{\mathrm{out},\theta}$ の場合も、しりとりの最初の単語が θ から始まるときに、出次数が先に 0 となり θ で終了する可能性が生じる。よって、しりとりは次の式を満たす文字 θ で終了する。

$$k_{\mathrm{in},\theta} \geq k_{\mathrm{out},\theta}. \tag{1}$$

(1) 式を満たす文字を終了可能文字と呼び、満たさない文字を終了不能文字と呼ぶ。図 5 は Moby-Dick 辞書の各文字 θ の $(k_{\mathrm{in},\theta}, k_{\mathrm{out},\theta})$ をプロットしたものである。この辞書における終了可能文字は、黄色の領域にある 8 文字であることがわかる。



(a) $k_{\mathrm{in},\theta} < k_{\mathrm{out},\theta}$　(b) $k_{\mathrm{in},\theta} > k_{\mathrm{out},\theta}$

図 4: 終了可能性について

3　平均場近似

考えている辞書ネットワークの隣接行列を A とする。文字 θ, ϕ について、行列要素 $A_{\theta\phi}$ は θ で始まり ϕ で終わる単語の数である。しりとりで単語が選択されるごとに隣接行列と各文字の次数は変化する。l 個の単語が選択された時点での隣接行列を $A^{(l)}$、文字 θ の入次数と出次数を $k^{(l)}_{\mathrm{in},\theta}, k^{(l)}_{\mathrm{out},\theta}$ と表すと、$k^{(l)}_{\mathrm{in},\theta}, k^{(l)}_{\mathrm{out},\theta}$ はそれぞれ隣接行列の列和と行和に

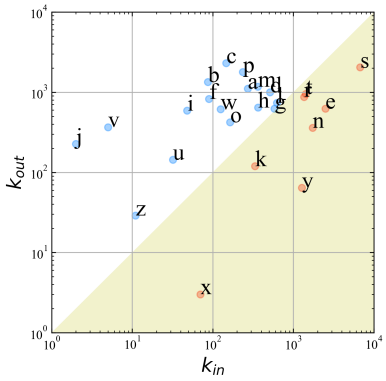


図 5: Moby-Dick 辞書の $k_{\mathrm{in}} - k_{\mathrm{out}}$ 平面（両対数）

なっており、

$$k^{(l)}_{\mathrm{in},\theta} = \sum_{\phi} A^{(l)}_{\phi\theta}, \quad k^{(l)}_{\mathrm{out},\theta} = \sum_{\phi} A^{(l)}_{\theta\phi} \tag{2}$$

と表される。

本研究ではしりとりの一回の試行を以下のように考える。(i) 初期条件として文字をランダムに選択する。すなわち、0 番目の単語の末尾文字はどの文字も $1/C$ の確率で選択される。(ii) l 番目の単語の末尾文字が θ のとき、$l+1$ 番目の単語の末尾文字に ϕ が選ばれる確率 $q^{(l)}_{\theta\phi}$ は以下の式で与えられる。

$$q^{(l)}_{\theta\phi} = \frac{A^{(l)}_{\theta\phi}}{\sum_{\phi'} A^{(l)}_{\theta\phi'}} = \frac{A^{(l)}_{\theta\phi}}{k^{(l)}_{\mathrm{out},\theta}} \tag{3}$$

ϕ で終わる単語が選択された後の隣接行列 $A^{(l+1)}$ は、$\theta\phi$ 成分のみが $A^{(l)}$ から 1 減少し、他の成分は変化しない。(iii) $k^{(l)}_{\mathrm{out},\theta} = 0$ となった後、θ で終わる単語が選ばれたらしりとりは終了し、そのときの単語列の長さ L を鎖長とする。

しりとりが鎖長 L かつ文字 θ で終了する確率を $p_\theta(L)$ で表す。一般の辞書に対して $p_\theta(L)$ を解析的に求めることは容易ではない。以下では $p_\theta(L)$ そのものではなく、平均場近似を施した鎖長分布 $\tilde{p}_\theta(L)$ を取り扱う。平均場近似とは、末尾文字として θ の次に ϕ が選ばれる確率を、(3) 式の代わりにその分母分子をそれぞれ θ について和をとったもの、

$$\tilde{q}^{(l)}_{\theta\phi} = \frac{\sum_{\theta} A^{(l)}_{\theta\phi}}{\sum_{\theta}\sum_{\phi'} A^{(l)}_{\theta\phi'}} = \frac{k^{(l)}_{\mathrm{in},\phi}}{D^{(l)}} \tag{4}$$

で置き換えた近似である。l ステップ目における辞書の総単語数を $D^{(l)}$ とした。$\tilde{q}^{(l)}_{\theta\phi}$ は隣接行列の各成分ではなく、その列和すなわち $k^{(l)}_{\mathrm{in},\theta}$ にのみ依存する。つまり、平均場近似は与えられた辞書の各文字の入次数と出次数を保ったまま、頂点同士をランダムにつなぎかえてできる一連の辞書群（シャッフル辞書群）に対する平均を意味する。そして、この近似によってしりとりの一連の過程を確率論における非復元抽出型の壺モデルとして扱うことが可能になる。

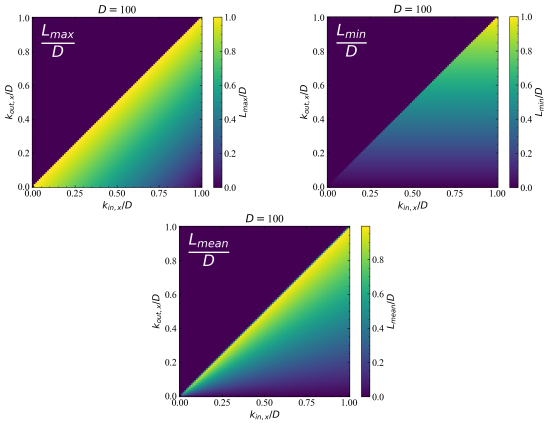


図 6: $\frac{L_{\max}}{D}, \frac{L_{\min}}{D}, \frac{L_{\mathrm{mean}}}{D}$ のヒートマップ ($D = 100$)

4 解析結果

最も簡単な例として、$C = 2$ のシャッフル辞書群における平均鎖長分布 $\tilde{p}_\theta(L)$ を求める。ただし、終了可能文字を x、終了不能文字を z とする。平均場近似において鎖長 L かつ文字 x で終了する確率は、x, z と書かれた球がそれぞれ $k_{\mathrm{in},x}$ 個, $k_{\mathrm{in},z}$ 個入っている壺の中から、非復元抽出で x の球を $k_{\mathrm{out},x}+1$ 個取り出した時点で、球が全部で L 個取り出されている確率と考えられる。よって、$\tilde{p}_x(L)$ は負の超幾何分布に従い、以下の式で表すことができる。

$$\begin{aligned}\tilde{p}_x(L) = &\frac{1}{2}\frac{\binom{L-1}{k_{\mathrm{out},x}-1}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}}}{\binom{D}{k_{\mathrm{in},x}}} \\ &+ \frac{1}{2}\frac{\binom{L-1}{k_{\mathrm{out},x}}\binom{D-L}{k_{\mathrm{in},x}-k_{\mathrm{out},x}-1}}{\binom{D}{k_{\mathrm{in},x}}}.\end{aligned} \tag{5}$$

L/D をカバーレートと呼ぶ。図 6 はカバーレートの最大値 $L_{\max}/D$、最小値 $L_{\min}/D$、平均値 L_{mean}/D を各点 $(k_{\mathrm{in},x}/D, k_{\mathrm{out},x}/D)$ ごとに (5) 式から求めたヒートマップである。この図から最小値は $k_{\mathrm{out},x}$ のみによって決まり、最大値と平均値はそれぞれ $k_{\mathrm{out},x}$ と $k_{\mathrm{in},x}$ の差と比によって決まると予想される。

(5) 式と負の超幾何分布の統計量 [7] をもとに、

$L_{\max}/D$、$L_{\min}/D$、L_{mean}/D を計算すると、

$$\frac{L_{\max}}{D} = 1 + \frac{k_{\mathrm{out},x} - k_{\mathrm{in},x} + 1}{D} \tag{6}$$

$$\frac{L_{\min}}{D} = \frac{k_{\mathrm{out},x}}{D} \tag{7}$$

$$\frac{L_{\mathrm{mean}}}{D} \simeq \frac{k_{\mathrm{out},x}}{k_{\mathrm{in},x}} \tag{8}$$

と求められ、予想が正しいことが確認された。ただし、平均値の計算には $D \gg 1$, $k_{\mathrm{in},x} \gg 1, k_{\mathrm{out},x} \gg 1$ の近似を用いた。数値計算でも (6)–(8) 式が成立することが確認された。図 7 は平均値についての計算結果の一例であり、D が大きくなるにつれて (8) 式の関係に漸近している。

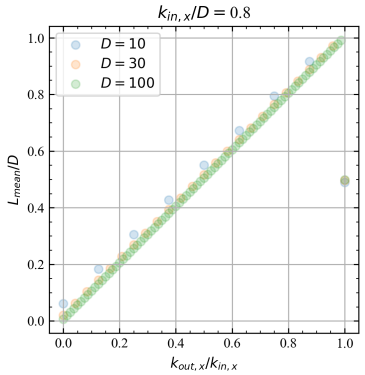


図 7: $\frac{L_{\mathrm{mean}}}{D}$ の $\frac{k_{\mathrm{out},x}}{k_{\mathrm{in},x}}$ 依存性 $\left(\frac{k_{\mathrm{in},x}}{D} = 0.8 \text{ に固定}\right)$

5　まとめと今後の課題

本研究では、ランダムなしりとりを辞書ネットワーク上の SAW とみなし、シャッフル辞書群におけるランダムなしりとりの平均鎖長分布を $C = 2$ の場合に限定して求めた。その結果、鎖長分布は負の超幾何分布に従うことが判明した。さらに、カバーレートの最小値は出次数のみで決まり、最大値と平均値はそれぞれ出次数と入次数の差と比で決まることが判明した。これらの統計量の次数依存性は数値計算によっても確認された。

なお $C = 2$ に限って言えば、シャッフル辞書群ではなく個別の辞書に対して鎖長分布を求めることが可能である。しかし、同じ手法で $C \geq 3$ の個別の辞書における鎖長分布を求めることはできていない。これに対して、本論文で紹介したシャッフル辞書群に対する平均場近似は $C \geq 3$ でも適用できるという利点を持っている。

$C \geq 3$ のシャッフル辞書群における平均鎖長分布は多変量化した負の超幾何分布となることが判明している。例えば、$C = 3$（文字 x, y, z のうち、x, y で終了可能）における $\tilde{p}_x(L)$ は以下のように表される。

$$\begin{aligned}
&\tilde{p}_x(L) \\
&= \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}} \frac{\binom{k_{\mathrm{in},x}}{k_{\mathrm{out},x}-1}\binom{k_{\mathrm{in},y}}{k_y}\binom{k_{\mathrm{in},z}}{L-k_{\mathrm{out},x}-k_y}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},x} - k_{\mathrm{out},x} + 1}{D - L + 1} \\
&+ \frac{1}{3}\sum_{k_y=1}^{k_{\mathrm{out},y}} \frac{\binom{k_{\mathrm{in},x}}{k_{\mathrm{out},x}}\binom{k_{\mathrm{in},y}}{k_y-1}\binom{k_{\mathrm{in},z}}{L-k_{\mathrm{out},x}-k_y}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},x} - k_{\mathrm{out},x}}{D - L + 1} \\
&+ \frac{1}{3}\sum_{k_y=0}^{k_{\mathrm{out},y}} \frac{\binom{k_{\mathrm{in},x}}{k_{\mathrm{out},x}}\binom{k_{\mathrm{in},y}}{k_y}\binom{k_{\mathrm{in},z}}{L-k_{\mathrm{out},x}-k_y-1}}{\binom{D}{L-1}} \times \frac{k_{\mathrm{in},x} - k_{\mathrm{out},x}}{D - L + 1}.
\end{aligned} \tag{9}$$

しかしその統計量は今のところ解析的に求められていない。今後は文字数が増えても、統計量は $C = 2$ と同じ性質を保つのか検証したいと考えている。

平均場近似の精度は単一辞書の鎖長分布とシャッフル辞書群の鎖長分布の距離で定量的に評価できる。Moby-Dick 辞書の場合、全変動距離の値は 0.160 であり、平均場近似はよい近似とは言えない。どのような辞書ならば平均場近似が妥当なのかという問題については、今後調査したいと考えている。

また、言語を変えると、各文字の $k_{\mathrm{in}} - k_{\mathrm{out}}$ の分布だけでなく、文字種類数 C も変化する場合がある。例えば日本語の場合 $C \geq 46$ となる。このように言語を変えたときにしりとりの統計的性質がどうなるのかも今後の興味深い話題である。

参考文献

[1] N. Inui, et al., Proceedings of the First International Conference on Informatics in Control, Automation and Robotics, **1** (2004) 214.

[2] M. Murata and T. Shirado, International Information Institute (Tokyo). Information, **18** (2015) 1631.

[3] Project Gutenberg, https://www.gutenberg.org/ebooks/2701.

[4] S. Hemmer and P. C. Hemmer, J. Chem. Phys., **81** (1984) 584.

[5] C. P. Herrero, Phys. Rev. E, **71** (2005) 016103.

[6] I. Tishby, et al., J. Phys. A: Math. Theor., **49** (2016) 285002.

[7] R. A. Khan, Sankhyā: The Indian Journal of Statistics, B(1960-2002), **56** (1994) 309.