第57回 プログラミング･シンポジウム 2016.1.8-10

正規言語の零壱則とその応用について

新屋 良磨[1,a)]

概要：「言語 L が零壱則に従う」とは，おおまかに言うと L が「ほとんど全ての文字列を含む」か「ほとんどどんな文字列も含まない」のどちらかを満たすことを言う．著者は [9] において正規言語 L が零壱則に従うことの代数的な必要十分条件，および線形時間での判定アルゴリズムを与えた．本論文では正規言語の零壱則とその代数的特徴付けについて説明し，そこから導かれる正規言語の非自明な性質，および零壱則の応用の可能性について議論を行う．具体的には，零壱則の「非正規性の判定」と「決定問題の近似」という 2 種類の応用について議論する．

キーワード：正規言語，零壱則，零元，無限の猿定理

1. はじめに

本論文では正規言語の零壱則の紹介とその応用について議論を行う．まず，正規言語の零壱則とは何か簡単に紹介してみよう（形式的な定義は次章にて行う）．「言語 L が零壱則に従う」とは，おおまかに言うと L が「ほとんど全ての文字列を含む」か「ほとんど全ての文字列を含まない」のどちらかを満たすことを言う．与えられた言語 L に対して「言語 L は全ての文字列を含むか？」を判定する問題は**充満性問題** (*universality problem*)，「言語 L はどんな文字列も含まないか？」を判定する問題は**空性判定** (*emptiness problem*) と呼ばれる．零壱則は充満性問題や空性問題のある種の近似と捉えることができる．言語 L が正規の場合，L が零壱則に従うことの代数的な必要十分条件は著者によって与えられており，さらに n 状態の決定性オートマトン $\mathcal{A}$ に対して $\mathcal{A}$ が受理する言語が零壱則に従うかどうかは線形時間 $\mathrm{O}(n)$ で判定することができる [9].

さて，本論文は本章を含め 6 章からなる．続く 2 章では零壱則の研究背景と有限文字列に対する零壱則の定式化を行う．正規言語に関する基本的な知識もこの章で紹介される．零壱則はもともと有限モデル理論という分野で研究されているものであり，有限グラフに関する一階述語理論の表現力の研究から生まれた．2 章では言語や論理式の例がいくつか出てくるが，形式言語理論や論理学の専門知識等はほとんど仮定していない．例題を見て感覚的に分かるように努めた．3 章では正規言語が零壱則に従う必要十分条件：**零壱定理**を紹介し，そのアイディアを解説する．4 章では零壱定理を非正規性 (与えられた言語が正規言語でないこと) の判定に応用する方法とその使用例を紹介する．5 章では先に述べた零壱則の「問題の近似」という点に着目し，その応用の可能性について議論を行う．最後の 6 章では正規言語の零壱則の今後の展望や未解決問題をまとめる．

[1] 東京工業大学 数理・計算科学専攻
[a)] shinya.r.aa@m.titech.ac.jp

2. 零壱則とは

零壱則 (*zero-one law*) は有限モデル理論の分野において深く研究されているある種の現象である．本章ではまず，古くから研究されている有限グラフの零壱則について解説を行い，その後，正規言語の零壱則 (文字列の零壱則) の定式化を行う．

2.1 有限グラフの零壱則

本論文では**グラフ**とは自己ループを含まない無向グラフを指す．グラフに対するある性質 P（例えば「平面グラフである」「連結である」など）と自然数 n について，以下の式を考えてみよう．

$$\frac{\text{性質 } P \text{ を満たす } n \text{ 頂点のグラフの総数}}{n \text{ 頂点のグラフの総数}} \tag{1}$$

式 (1) は「n 頂点のグラフからランダムに 1 つ取ってきたグラフが性質 P を満たす確率」を表している．グラフに対する性質 P が**ほとんど真** (*almost surely true*) であるとは，式 (1) で n に対する極限 $\lim_{n\to\infty}$ が 1 に収束すること言う．逆に，P が**ほとんど偽** (*almost surely false*) であるとは，式 (1) で n に対する極限が 0 に収束することを言う．ある性質 P がほとんど真であるとは，感覚的には「P がほとんど全ての有限グラフで成り立つ」と捉えることができる．例えば以下のグラフの性質はほとんど真である ([6])．

- 連結 (任意の 2 頂点間にパスが存在する)
- ハミルトニアン閉路が存在する
- 3 彩色可能でない
- 非自明な自己同型が存在しない (*rigid*).

グラフに対するある性質 P を固定すれば，P がほとんど真かほとんど偽かどうかはただの組合せ的な問題である．一方，性質 P がどの程度の強さの論理で表現できるかは論理学の問題である．ある論理 $\mathcal{L}$ が **(有限グラフにおいて) 零壱則を持つ** (*$\mathcal{L}$ has the zero-one law*) とは，$\mathcal{L}$ で表現できる**任意のグラフの性質**が「ほとんど真」か「ほとんど偽」のいずれかしか無いことを言う．例えば，**一階述語論理** (*first-order logic*) で定義可能なグラフの全ての性質はほとんど真かほとんど偽のいずれかしか存在しないことが知られている．

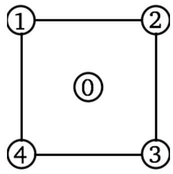

図 1 孤立点を持つグラフ

定理 1 (Glebskiietal et al. [5], Fagin [3])**.** グラフに対する一階述語論理は零壱則を持つ．

例題 1. 次の一階述語の論理式 Φ を考えてみよう．

$$\Phi = \exists x\Big(\forall y\big(\neg E(x,y)\big)\Big).$$

なお，変数 x, y はグラフの頂点を，述語 $E(x,y)$ は 2 頂点 x, y 間に辺が存在するかどうかを表している．この Φ はすなわち「どの頂点にも繋がっていない頂点 x(**孤立点**) が存在する」ことを表現した一階述語論理式である．例えば図 1 のグラフは Φ を満たす (0 が孤立点である)．n 頂点のグラフが孤立点を持つ確率は $\frac{n}{2^{(n-1)}}$ 以下である (孤立点 x の取り方 n 通りに，自身を除く $n-1$ 頂点全てと繋がっていない確率 $2^{-(n-1)}$ を掛けたもの) ことが分かる．そのため Φ はほとんど偽となる．

定理 1 は (例 1 の Φ だけでなく) **任意の**一階述語論理で定義可能な性質がほとんど偽かほとんど真のいずれかしかないことを言っている．1976 年に Fagin は**ランダムグラフ**と呼ばれるある特殊な (無限頂点の) グラフを用いて次の定理を証明し，その系として定理 1 を導いた．

定理 2 (Fagin [3])**.** 一階述語論理式 Φ がほとんど真である $\Leftrightarrow$ Φ がランダムグラフで成り立つ[*1]．

ある論理式 Φ がほとんど真であるかどうかは，素朴には無限個の有限グラフについて真偽を考える必要がある．しかし，定理 (2) はたった 1 つのグラフについて真偽を考えれば良いことを述べている美しい定理である．一階述語論理式のランダムグラフに対する真偽判定は決定可能であることが知られているため，定理 (2) は与えられた一階述語論理式が「ほとんど真 (ほとんど偽)」かどうかの判定問題の決定可能性も導く (*cf.* [8] の 12 章)．

[*1] $\Leftrightarrow$ は左辺と右辺が等価な命題 (必要十分条件) であることを表す

2.2 文字列の基礎と正規言語

さて，有限グラフの零壱則の次は本論文のメインとなる**有限文字列の零壱則**に解説を進めたいのだが，その前に文字列に関する基礎的なことをおさらいしておこう．

本論文では**文字** (*letter*) の有限集合を A で表す (*Alphabet* の A)．A 上の**文字列** (*string*) とは A に属する文字 a_i を有限個並べた列

$$a_0a_1\cdots a_n$$

である．文字列には**連接** (*concatenation*) という演算を考えることができ，2 つの文字列 $u = a_0a_1\cdots a_i, v = b_0b_1\cdots b_j$ の連結 uv は u と v を並べたものを表す：

$$uv = a_0a_1\cdots a_ib_0b_1\cdots b_j.$$

文字列 $w = a_0a_1\cdots a_n$ の**長さ**を $|w|$ で表す：$|w| = n+1$．2 つの文字列 u, v を連接した文字列の長さは自明に $|uv| = |u| + |v|$ を満たす．ε という記号で長さが 0 の文字列 (**空文字列**) を表す．空文字列は連接という演算に対して単位元となる．すなわち任意の文字列 w に対して：

$$\varepsilon w = w\varepsilon = w.$$

A に属する文字から作られる文字列全ての集合を A^* で表す．A に属する文字から作られる長さ n の文字列全ての集合を A^n で表す．例えば $A = \{0,1\}$ の場合は

$$A^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \cdots\}$$

となり，これは任意のバイナリ列を表す (注：どんな文字集合 A に対しても A^* は常に空文字列 ε を含む約束にしておく)．文字列 $w = a_0a_1\cdots a_n$ に対し，その**反転** (*reverse*) を $w^r = a_na_{n-1}\cdots a_0$ で表す．例えば，$abbab$ の反転は $(abbab)^r = babba$ となる．文字列 u, v に対し，ある文字列 x, y が存在して $u = xvy$ となる時，v を u の**部分文字列** (*factor*) と呼ぶ．

文字集合 A 上の**言語** (*language*) とは A^* の部分集合のことを指す．すなわち $L \subseteq A^*$ となる L を言語と呼ぶ．言語 L に対して L の要素の数を $|L|$ で表す．なお A^* は無限集合 $|A^*| = \infty$ である．

注 1. 一般に言語と言うと「日本語」や「フランス語」などの自然言語を連想するかもしれないが，形式言語理論・計算理論という研究分野においては言語を「文字列の集合」という身も蓋もない素朴な対象として扱う．

文字 A 上の 2 つの言語 L, K に対して次の 3 つの演算を定義する：

(集合) 和 $\quad L \cup K = \{w \in A^* \mid w \in L \vee w \in K\},$
連接 $\quad LK = \{uv \in A^* \mid u \in L \wedge v \in K\},$
Kleene 閉包 $\quad L^* = \{\varepsilon\} \cup L \cup LL \cup \cdots = \bigcup_{n\geq 0} L^n.$

ここで L^n は言語 L の n 回の連接を表し，L^0 は $\{\varepsilon\}$ で定義される．言語 L の Kleene 閉包 L^* は，直感的には「L に属する文字列を連接して作ることのできる任意の文字列集合」と捉えることができる (先に定義した A^* は正に A の Kleene 閉包である)．**正規言語** (*regular language*) とは，有限要素の言語から上記 3 つの演算の有限回の組合せによって構成できる言語である．

定義 1. 文字集合 A について：

- 空集合は正規言語である．
- 任意の文字 $a \in A$ についてその単集合 $\{a\}$ は正規言語である．

さらに，L と K が正規言語であるならばその和 $L \cup K$，連接 LK，Kleene 閉包 L^* はいずれも正規言語である．上記の形で導出されるもののみが正規言語である．

単集合な言語 $\{a\}$ は単に a と書く場合もある．また言語 L の**補集合** (*complement*) を $\overline{L}$ で表す．

例題 2. 例えば，文字集合 $A = \{a, b\}$ について，以下の言語は全て正規言語である．

- 偶数長の文字列の集合：
  $\{w \in A^* \mid |w|$ は偶数 $\} = (AA)^*.$
- 文字 a が最初にくる文字列の集合：
  $\{aw \mid w \in A^*\} = aA^*.$
- 文字 a の任意回の繰り返しの後に b の任意回の繰り返しが来る文字列の集合：
  $\{a^nb^m \mid n, m \geq 0\} = a^*b^*.$
- 文字列 v を部分文字列に含む文字列の集合：

$\{uvw \mid u, w \in A^*\} = A^* w A^*.$

例題 3. 一方，次の言語はいずれも正規言語では**ない**．証明は 4 章で行う．

- 偶数長の**回文**の集合：$\{ww^r \mid w \in A^*\}$.
- 文字集合 $A = \{[,]\}$ 上の **Dyck 言語** (角括弧のバランスが取れた文字列の集合，形式的な定義は 4 章で行う)：
  $\{\varepsilon, [], [[]], [][], [[[]]], [[][]], [[]][], [][[]], [][][], \ldots\}$.

正規言語という概念は 1951 年に Kleene[7] によって生み出されされ，以来，理論的に非常に良い性質を持つことが明らかにされている．プログラミングにおいては文字列パターンを記述するための道具：**正規表現** (*regular expression*) として重宝されている．正規言語の歴史やプログラミングにおける用法については [12] を参照されたし．

2.3 有限文字列の零壱則

論理的に表現できる性質というものは何もグラフの性質に限るものではない．理論計算機科学においても，実世界のプログラミングにおいても，**文字列に対する性質**を表現・検査することはもっとも基本的なことである．

グラフに対する論理では変数は頂点を表し，「2 頂点 x, y 間に辺が存在するか」という (2 引数の) 述語 $E(x, y)$ が使えた．一方，文字列に対する論理では変数は**文字列の添字** (*index*) を表し，各文字 a に対し「i 番目の文字は a である」という (1 引数の) 述語 $P_a(i)$ が使える．例えば，$aabb$ という文字列に対し $P_a(0)$ と $P_a(1)$ は真であり，$P_b(2)$ と $P_b(3)$ もまた真である (逆に $P_a(2), P_a(3), P_b(0), P_b(1)$ は全て偽である)．

例題 4. 次の一階述語論理式 $\Phi_{a^*b^*}$ を見てみよう．

$$\Phi_{a^*b^*} = \neg\Big(\exists i, j\big((i < j) \land P_b(i) \land P_a(j)\big)\Big).$$

この論理式 $\Phi_{a^*b^*}$ は「2 つの添字 $i < j$ で $P_b(i)$ かつ $P_a(j)$ となるようなものは存在しない」，つまり **b の後には a が出現しない**という性質を表している．$\Phi_{a^*b^*}$ を真にするような文字列の集合を $L(\Phi_{a^*b^*})$ で表記すると次のような無限集合となる：

$L(\Phi_{a^*b^*}) = \{\varepsilon, a, b, aa, ab, bb, aaa, aab, abb, \cdots\}$.

すなわち論理式 $\Phi_{a^*b^*}$ は a^*b^* という正規言語を一階述語論理で記述したものである．

1960 年に Büchi[2] は正規言語の論理的な特徴付けを与えた．

定理 3 (Büchi [2])**.** 言語 L が正規である $\Leftrightarrow$ L が**単項二階述語論理** *(monadic second-order logic)* で定義可能．

詳細は説明しないが，単項二階述語論理とは一階述語論理よりも真に表現力の高い論理である (*cf.* [8] の 7 章)．Büchi による定理 3 の発見以来，一階述語論理やその他様々な論理が正規言語のサブクラスに対応することが明らかにされてきた．さて，前節ではグラフ上の論理 $\mathcal{L}$ に対して「$\mathcal{L}$ が零壱則を持つ」という概念を定義した．文字列上の論理に対しても類推によって零壱則の定義が可能なことが読み取れるであろうが，まず「言語 L が零壱則に従う」という概念を定義しよう．

文字集合 A 上の言語 L に対して，L の**数え上げ関数** $\gamma_n(L)$ は自然数 n に対して L に属する長さ n の文字列の総数を表す：

$$\gamma_n(L) = |\{w \in L \mid |w| = n\}| = |L \cap A^n|.$$

さらに，L の**確率関数** $\mu_n(L)$ を自然数 n に対して

$$\mu_n(L) = \frac{|L \cap A^n|}{|A^n|} = \frac{|L \cap A^n|}{|A|^n}.$$

と定義する．$\mu_n(L)$ は「A^n からランダムに取ってきた文字列が L に属する確率」に他ならない．L の確率関数の n に対する極限

$$\lim_{n\to\infty} \mu_n(L)$$

が存在する時，それを **L の測度**と呼び $\mu(L)$ で表す[*2]．

定義 2. 文字集合 A 上の言語 L が**ほとんど充満** (*almost full*) であるとは $\mu_n(L) = 1$ が成り立つことを言う．逆に，言語 L が**ほとんど空** (*almost empty*) であるとは $\mu(L) = 0$ が成り立つことを言う．言語 L が**零壱則に従う** (*obeys the zero-one law*) とは，L がほとんど充満かほとんど空のいずれかであることを言う．

[*2] 厳密には μ は測度論で言う所の測度の性質は満たさないが，ここでは測度と呼んでいる

定義 2 は正規言語に限らず一般の言語に対して定義されていることに注意．

文字列上の論理 $\mathcal{L}$ が零壱則を持つということは，すなわち $\mathcal{L}$ で定義できる任意の言語が零壱則に従うことに他ならない．ところで，グラフ上の論理に対しては一階述語論理が零壱則を持つ (定理 1) ことを前節で述べたが，文字列上の論理に対してはそのような結果は成り立つのだろうか？ 実は，文字列上の一階述語論理は零壱則を持たない．すなわち，一階述語論理で定義可能な言語で測度が 0 でも 1 でもない言語が存在する．

例題 5. 次の一階述語論理式 Φ_{aA^*} について考えてみる．

$$\Phi_{aA^*} = \exists i\bigl(P_a(i) \land \forall j(i \leq j)\bigr).$$

Φ_{aA^*} は「最初の文字が a である文字列」，すなわち正規言語 aA^* を表現していることが容易に確かめられる．しかし任意の n に対して aA^* の確率関数は $\mu_n(aA^*) = 1/|A|$(最初の 1 文字が a である確率) を満たし，よってその極限である aA^* の測度も $1/|A|$ となる．一般に A が 2 つ以上の文字を含む ($|A| \geq 2$) 場合において aA^* は測度が 0 でも 1 でもない．そのため一階述語論理は零壱則を持たない．

ここで，次のような問題が自然に湧いてくる．

問題 1. どのような文字列上の論理が零壱則を持つのだろうか？

問題 2 (正規言語の零壱則)**.** 正規言語 L が零壱則に従う決定可能な必要十分条件は存在するだろうか？

3. 正規言語の零壱則

ある文字列 $w \in A^*$ に対して A^*wA^* と表現される言語を (w で生成された A 上の) **イデアル言語** (*ideal language*) と呼ぶ．著者は [9] にて問題 2 に対する肯定的な次の答えを与えた．

定理 4 (**零壱定理**, S. [9])**.** 文字 A 上の正規言語 L について，以下の条件は等しい：

(1) L が零壱則に従う

(2) L かその補集合 $\overline{L}$ がイデアル言語を含む．

より厳密には「正規言語 L が零壱則に従う」ということと「L の *syntactic monoid* が零元を持つ」「L が零オートマトンという特殊なオートマトンで認識できる」という条件が等しいことを示し，オートマトン的な特徴付けから線形時間の判定アルゴリズムを構成した．詳細は [9] を参照されたし．

定理 5 (S. [9])**.** 正規言語 L が零壱則に従うかどうかを判定する線形時間アルゴリズムが存在する．ただし，L は決定性オートマトンで与えられるものとし，線形時間とはその状態数に対して線形時間であることを意味する．

例題 6. 零壱則に従わ**ない**正規言語を例に，零壱定理の主張を確かめてみよう．

- 正規言語 $(AA)^*$ は零壱則に従わない．なぜなら：

$$\mu_n((AA)^*) = \begin{cases} 1 & n \text{ が偶数の場合,} \\ 0 & n \text{ が奇数の場合,} \end{cases}$$

  が成り立つため，その極限 $\mu((AA)^*)$ は存在しない *(0 と 1 で振動する)*．$(AA)^*$ は A^*wA^* という形の言語を含まないことは明らか．

- 例題 *5* で述べたように，正規言語 aA^* は零壱則に従わない．aA^* が A^*wA^* という形の言語を含まないことは

本論文では定理 4 と定理 5 の証明の詳細には踏み込まない．しかし，定理の直感を伝えるために零壱定理の簡単な方向 (2) ⇒ (1) の解説を行う．実は (2) ⇒ (1) の証明は非常に容易である．実際，次の良く知られた**無限の猿定理**から直ちに導かれる．

定理 6 (**無限の猿定理**)**.** 任意の文字列 $w \in A^*$ について，十分長い n を取れば A^n からランダムに取ってきた文字列は w を**ほとんど確実に**部分文字列として含む．形式的に言えば $\mu(A^*wA^*) = 1$ が成り立つ．

もちろん，無限の猿定理の証明は容易であるが，本論文では割愛する．さて，無限の猿定理を用いて実際に (2) ⇒ (1) を示そう．

命題 1. A 上の言語 L に対して，L かその補集合 $\overline{L}$ がイデアル言語を含むならば，L は零壱則に従う．

Proof. L が A^*wA^* という言語を含む，すなわ

ち $A^*wA^* \subseteq L$ と仮定する．無限の猿定理より A^*wA^* の測度は 1 である．測度の定義より明らかに $\mu(L) \geq \mu(A^*wA^*)$ が成り立つため，$\mu(L) = 1$ となる．すなわち，L が A^*wA^* を含む場合は L はほとんど充満となる．同様に，$\overline{L}$ が A^*wA^* を含む場合は L はほとんど空となることが導ける． □

実際には命題 1 は言語 L が正規であることを仮定していないので，零壱定理の (2) ⇒ (1) よりも強い主張である．零壱定理の本質は言語 L が正規の場合に (1) ⇒ (2) が成り立つという部分であり，その証明は容易ではない．[9] では零壱則に従う言語全体の閉包性と **Eilenberg の補題**と呼ばれる正規言語の構造定理を用いて (1) ⇒ (2) を示した．

零壱定理における条件 (1) と (2) の等価性は正規言語において成り立つものであり，一般の言語については成り立たない．その点を次章で明らかにする．

4. 零壱則による非正規性の証明

命題 1 より，（正規に限らない）言語 L について，L かその補集合 $\overline{L}$ がイデアル言語を含む場合 L は零壱則に従う．零壱定理より，言語 L が正規の場合はその逆も成り立つ．しかし，言語 L が非正規の場合はその逆は必ずしも成り立たない．もし，零壱則に従うが，L も $\overline{L}$ もイデアル言語を含ま**ない**言語 L が存在すれば，その L は零壱定理より**非正規である**．これが零壱定理を使った非正規性の証明のアイディアである．実際に本章では零壱定理を使って例題 3 で紹介した 2 つの言語「偶数長の回文の集合」と「Dyck 言語」が非正規であることを示す．

4.1 非正規性のための零壱補題

「L か $\overline{L}$ がイデアル言語を含む」という条件をより分かりやすく整形し，零壱定理を非正規性の証明ために変形したものが次の補題である．

補題 1 (**零壱補題**)**.** 零壱則を満たす A 上の言語を L とする．もし L が空でも充満でもなく ($L \neq \emptyset \wedge L \neq A^*$)，さらに次の条件 (**非零条件**) を満たすならば，L は非正規：

$$\forall v \in A^* \Big(\exists u, w \in A^* \, (v \in L \Leftrightarrow uvw \notin L)\Big).$$

Proof. 非零条件の否定を取ると次のようになる：

$$\exists v \in A^* \Big(\forall u, w \in A^* (v \in L \Leftrightarrow uvw \in L)\Big).$$

言語 L がこの否定の条件を満たすということは，論理式の形から明らかなように，L か $\overline{L}$ がイデアル言語 A^*vA^* を含むことになる．A^* でも $\emptyset$ でもない言語 L が非零条件を満たすならば，**L も $\overline{L}$ もイデアル言語を含まない**ことになる．一方，L は零壱則を満たす．零壱定理より，もし L が正規言語であるならば L か $\overline{L}$ がイデアル言語を含むことになる．これは矛盾である．よって L は非正規な言語となる． □

4.2 偶数長の回分の集合の非正規性

文字 A における偶数長の回分の集合を P とおく：

$$P = \{ww^r \mid w \in A^*\}.$$

もし A が 1 つの文字 a しか含まない場合 ($A = \{a\}$) は，$P = (aa)^*$ となり，正規言語になることに注意されたい．零壱補題を用いて $|A| \geq 2$ の場合での非正規性を示そう．

命題 2. *A が少なくとも 2 つの文字を含む ($|A| \geq 2$) とき，P は非正規．*

Proof. P に属する文字列は必ず ww^r の形をしている．つまり，P に属する長さ $2n$ の文字列は長さ n の前半部分 (*prefix*) によって一意に定まる．よって P の確率関数は次の等式を満たす：

$$\mu_n(P) = \begin{cases} \frac{|A|^{n/2}}{|A|^n} = \frac{1}{|A|^{n/2}} & n \text{ が偶数の場合,} \\ 0 & n \text{ が奇数の場合.} \end{cases}$$

そのため，$\mu_n(P)$ は 0 に収束する．つまり，P はほとんど空となる．

さて，P が零壱補題の非零条件を満たすことを確認する．これはとても容易である．実際，P に属する任意の文字列 w について，文字 $a \in A$ を連接した文字列 wa は長さが奇数になるため明らかに P に属さない．逆に，P に属さない任意の文字列 w について，w にその反転を連接した文字列 ww^r は P に属する．よって P は非零条件を満たし，零壱補題より P は非正規． □

4.3 Dyck 言語の非正規性

例題 3 では $A = \{[,]\}$ 上の Dyck 言語 D を「角括弧のバランスが取れた文字列」と定義した：

$$D = \{\varepsilon, [], [[]], [][], [[[]]], [[][]], [[]][], [][[]], [][][], \ldots\}.$$

本節ではより形式的な D の定義を与える．w を $A = \{[,]\}$ 上の文字列とする．ここで関数 $\mathsf{Trim} : A^* \to A^*$ を「文字列 w から w 中の [] という形の部分文字列を全て消去したもの」と定義する．例えば，[], [[] と [[]]][][はそれぞれ次のように Trim に二重下線部の $\underline{\underline{[]}}$ が消去される：

$$\mathsf{Trim}(\underline{\underline{[]}}) = \varepsilon, \mathsf{Trim}([\underline{\underline{[]}}) = [, \mathsf{Trim}([\underline{\underline{[]}}]]\underline{\underline{[]}}[) = []][.$$

Trim の定義から明らかなように，任意の文字列 w に対して w **から始まる Trim の不動点**が存在する．つまり，ある自然数 $n \geq 1$ が存在し，任意の自然数 $m \geq n$ に対して $\mathsf{Trim}^n(w) = \mathsf{Trim}^m(w)$ が成り立つ．この時 $\mathsf{Trim}^m(w)$ を w **の既約文字列**と呼ぶ．任意の文字列 w に対してその既約文字列を $\mathsf{Trim}^* w$ で表す．すると Dyck 言語 D は次のように定式化することができる：

$$D = \{w \in A^* \mid \mathsf{Trim}^*(w) = \varepsilon\}.$$

さて，零壱補題を用いて D の非正規性を示そう．

命題 3. $A = \{[,]\}$ 上の *Dyck* 言語 D は非正規．

Proof. 組合せ理論の初等的な事実を用いる．自然数 $n \geq 1$ に対して，$\gamma_{2n}(D)$ の値：

$$\gamma_2(D) = 1, \gamma_4(D) = 2, \gamma_6(D) = 5, \gamma_8(D) = 14, \cdots$$

は n 番目の**カタラン数** (*Catalan number*) となることが知られている．また，カタラン数は n に対して $\Theta(\frac{2^n}{n^{3/2}})$ に漸近することが知られている (*cf.* [4])．カタラン数の事実を用いると，D の確率関数は次のようになる：

$$\mu_n(D) = \begin{cases} \Theta\left(\frac{1}{n^{3/2}}\right) & n \text{ が偶数の場合,} \\ 0 & n \text{ が奇数の場合.} \end{cases}$$

よってその極限は存在し 0 に収束する．つまり，D はほとんど空となる．

次に，D が零壱補題の非零条件を満たすことを確認する．D に属する任意の文字列 w について，$w[$ は明らかに D に属さない（長さ $|w[|$ が奇数となるし，そもそも角括弧のバランスが崩れる）．逆に，D に属さない任意の文字列 w について考える．D の定義より w の既約文字列は空文字列ではなく，明らかに：

$$\mathsf{Trim}^*(w) =]^n[^m$$

という形をしている．ここで n, m は自然数でどちらか一方は 0 でない．この時，w に左から $[^n$，右から $]^m$ を連接した文字列 $[^n w]^m$ は D に属する．なぜなら Trim の定義より次の等式が成り立つからである：

$$\mathsf{Trim}^*([^n w]^m) = \mathsf{Trim}^*([^n]^n[^m]^m) = \varepsilon.$$

よって D は非零条件を満たし，零壱補題より D は非正規. □

5. 零壱則による問題の近似

与えられた言語 L に対して「L は充満 ($L = A^*$) か？」を判定する問題を充満性問題，「L は空か ($L = \emptyset$)？」を判定する問題は**空性問題**と呼ばれた．「L は**ほとんど**充満か？」「L は**ほとんど**空か？」を考える零壱則の観点は，充満性問題と空性問題の**近似**的な問題を与える．

厳密に解こうとすると計算不能・計算量が高いような問題も，その近似問題を考えることで計算可能・計算量が低くなる場合がある．実際，有限グラフの世界には美しく輝く “零壱則による近似の成功例” が存在する．

事実 1. 与えられた一階述語論理式 Φ について：

- Φ が**全ての**有限グラフで成り立つかどうか (有限恒真性判定) は**決定不能**.
- Φ がほとんど全ての有限グラフで成り立つかどうかは**決定可能**.

Φ の充足可能性問題の決定不能性は 1950 年に Trakhtenbrot が示した有限モデル理論の有名な定理である [11]．Φ がほとんど真であることの決定可能性は Fagin の定理による．

事実 1 のような有限グラフの世界における零壱

則による近似の成功例が，有限文字列の世界にも存在しないだろうか？本論文では正規言語の零壱則の決定可能な特徴付けを述べたが，しかし，正規言語は元々非常に扱いやすい言語クラスである．多くの判定問題 (例えば充満性や空性など) は決定可能であり，事実 1 のような成功例は望めないだろう．しかし，正規言語より上位の言語クラスに対する零壱則の近似を考えることには意義があるように思える．例えば**文脈自由言語** (*context-free language*) と呼ばれる言語は一般に**充満性問題が決定不能**であることが広く知られている (*cf.* [10])．

問題 3. 文脈自由言語 L が零壱則に従うかどうかは決定可能だろうか？

正規言語の零壱則の証明の要点は「零壱則を満たす言語族の閉包性」と「正規言語の構造定理 (Eilenberg の補題)」[9] であり，「零壱則を満たす言語族の閉包性」は正規言語に限らない一般の言語クラスでの議論であり，拡張の可能性がある．正規言語の零壱則の成果を一気に文脈自由言語に拡張するのは無謀な話ではあるが，正規言語より強く，文脈自由言語より弱い言語クラスについて少しずつ零壱則の研究を進めていくアプローチは可能性がある (例えば **visibly pushdown language (VPL)**[1] や**決定性文脈自由言語** (*cf.* [10]) など)．

6. おわりに

本論文では正規言語の零壱則という概念と零壱定理の紹介を行い，2 種類の応用の可能性として

(1) 零壱補題 (補題 1) による非正規性の証明

(2) 零壱則による近似の応用可能性

の議論を行った．

(1) については，零壱補題を用いた「偶数長の回文の集合」と「Dyck 言語」の非正規性の簡潔な証明を与えた．言語の非正規性を示す最も有用な道具の一つに**ポンピング補題** (*pumping lemma*) が挙げられる．零壱補題とポンピング補題は「どちらが便利である」などと比べられるものではないが，(両者の内容を理解している読者であれば) それぞれ全く異なる技法を使って非正規性を証明することが分かるだろう．零壱補題による非正規性の証明は大きく 2 つの行程に分けられる：「確率関数の解析」と「非零条件の考察」．言語に対する数え上げ関数は純粋に組合せ論的な対象であり，多くの言語において数え上げ関数の一般項や計算量が明らかになっている場合は多い (Dyck 言語の場合のカタラン数のように)．(1) の今後の課題として，ポンピング補題では証明が難しく (あるいは不可能)，零壱補題では証明が容易であるような自然な非正規言語の具体例を探すことが考えられる．

(2) については，あくまで可能性を示すだけのものであり，具体的な応用は提案していない．応用の嗅覚に長けた方々にもぜひ考えてもらいたい．正規言語の零壱則の基礎づけ [9] が，より上位の言語クラスへ零壱則の理解を深めるための土台になれば嬉しい．

謝辞　論文の校正をしてくれた加藤肇君 (東京工業大学) に感謝を．本研究は JSPS 科研費 14J11962 の助成を受けたものである．

参考文献

[1] Alur, R., Kumar, V., Madhusudan, P. and Viswanathan, M.: Congruences for Visibly Pushdown Languages, *Automata, Languages and Programming* (Caires, L., Italiano, G., Monteiro, L., Palamidessi, C. and Yung, M., eds.), Lecture Notes in Computer Science, Vol. 3580, Springer Berlin Heidelberg, pp. 1102–1114 (2005).

[2] Büchi, J. R.: Weak second-order arithmetic and finite automata, *Z. Math. Logik und Grundl. Math.*, Vol. 6, pp. 66–92 (1960).

[3] Fagin, R.: Probabilities on Finite Models, *J. Symb. Log.*, Vol. 41, No. 1, pp. 50–58 (1976).

[4] Flajolet, P. and Sedgewick, R.: *Analytic Combinatorics*, Cambridge University Press, New York, NY, USA, 1 edition (2009).

[5] Glebskii, Y. V., Kogan, D. I., Liogonkii, M. I. and Talanov, V. A.: Range and degree of realizability of formulas in the restricted predicate calculus., *Cybernetics*, Vol. 5, pp. 142–154 (1969).

[6] Gurevich, Y.: Zero-One Laws, *Bulletin of the EATCS*, Vol. 46, pp. 90–106 (1992).

[7] Kleene, S. C.: *Representation of events in nerve nets and finite automata*, Rand Corporatin (1951).

[8] Libkin, L.: *Elements of Finite Model Theory*, SpringerVerlag (2004).

[9] Sin'ya, R.: An Automata Theoretic Approach to the Zero-One Law for Regular Languages: Algorithmic and Logical Aspects, *Proceedings Sixth International Symposium on Games, Automata, Logics and Formal Verification, GandALF 2015*, pp. 172–185 (2015).

[10] Sipser, M.: *Introduction to the theory of computation: third edition*, Cengage Learning, Boston, MA, USA, 3 edition (2012).

[11] Trakhtenbrot, B.: The impossibilty of an algorithm for the decision problem for finite models (in Russian), *Doklady Akademii Nauk SSR*, Vol. 70, pp. 569–572 (1950).

[12] 新屋良磨. 鈴木勇介. 高田謙：正規表現技術入門, 技術評論社 (2015).