正則言語で極限的に近似可能な言語について

新屋　良磨 [1]

[1] 秋田大学　数理科学コース

ryoma@math.akita-u.ac.jp

概 要　著者は [21] において正則可測性という概念を導入し，多くの複雑な文脈自由言語が正則可測である一方,「原始語全体の集合」と言う組合せ論的に重要な言語が強い意味で非可測であることを示した．ある言語 L が正則可測であるとは，直感的には L が正則言語でいくらでも近似できる，すなわち L に「収束」する正則言語の無限列が存在することを言う．

本論文では [21] で考察されていなかった，可測な言語全体の一般的な性質について議論する．また，正則言語の部分族であるいくつかの言語族について，その言語族において可測な言語全体の集合を明らかにする．

1　はじめに

非常に複雑な形をしたオブジェの体積を測るにはどうすれば良いだろうか？もしそれが水に濡らしても良いものであれば，手っ取り早い方法は水を満たした直方体の桶に糸で吊ったオブジェをゆっくりかつ完全に沈め，引き出し，減った水位から溢れた水の量を計算すれば良い．溢れた水の量はオブジェを「覆う」のに必要な水の量なので，それがオブジェの体積の良い目安になるだろう．このように，図りたい対象 $X \subseteq \mathbb{R}^d$ を基本集合と呼ばれる性質の良い集合 $Y \supseteq X$ で覆い，Y の体積を X の体積の目安 (外側からの近似値) とすることは測度論においても常套手段である．

例えばLebesgue測度においては，実数 $a \leq b \in \mathbb{R}$ に対し区間 $I = [a,b], [a,b), (a,b], (a,b)$ の長さを $|I| = b - a$ で定め，d 個の区間の直積 $B = I_1 \times \ldots \times I_d$ を $\mathbb{R}^d$ の直方体 (体積は $|B| = |I_1| \times \cdots \times |I_d|$) と呼び，**直方体の可算和を基本集合とみなす**．集合 $X \subseteq \mathbb{R}^d$ の Lebesgue 外測度は

$$m^*(X) = \inf \left\{ \sum_{n=1}^{\infty} |B_n| \;\middle|\; \bigcup_{n=1}^{\infty} B_n \supseteq X; \text{各 } n \text{ で } B_n \text{は直方体} \right\}$$

として定義され，これは X を基本集合で「覆う」のに必要な体積の下限となっている．X が Lebesgue 可測であるとは，

$$\forall S \subseteq \mathbb{R}^d \qquad m^*(S) = m^*(S \cap X) + m^*(S \cap \overline{X})$$

を満たすことであった．上の条件は Carathéodory 条件と呼ばれる．

自然数全体 $\mathbb{N}(\ni 0)$ のような可算無限な集合に対しても，実はこの測度論的な手法は適用できる．Buck [5] はまず自然数 $p, q \in \mathbb{N}$ に対して等差数列 $A = \{pn + q \mid n \in \mathbb{N}\}$ [1]の密度を $d(A) = 1/p$ (ただし $p = 0$ の場合は $d(A) = 0$) と定め，**等差数列の有限和を基本集合とみなし**，$X \subseteq \mathbb{N}$ の外密度を

$$d^*(X) = \inf \left\{ \sum_{n=1}^{k} d(A_n) \;\middle|\; \bigcup_{n=1}^{k} A_n \supseteq X; k \in \mathbb{N} \text{ かつ各 } n \text{ で } A_n \text{は等差数列} \right\}$$

で定義した．Lebesgue 測度のときと同様に，Buck は $X \subseteq \mathbb{N}$ が Carathéodory 条件

$$\forall S \subseteq \mathbb{N} \qquad d^*(S) = d^*(S \cap X) + d^*(S \cap \overline{X})$$

を満たすときに X は可測と言い，値 $d^*(X)$ を X の測密度 (*measure density*) と呼んだ．

[1] p は 0 であっても良く，単元集合 $\{q\}$ も等差数列と本論文では呼ぶ．

著者が [21] で導入した正則可測性は，Buck [5] の測密度の形式言語への拡張となっている．すなわち，言語 $L \subseteq A^*$ に対して密度を定義し，**正則言語を基本集合とみなし**，外密度を経由して言語の可測性を定義する (厳密な定義は次節で行う)．$A = \{a\}$ の場合は A^* と $\mathbb{N}$ は同一視でき，この場合は正則可測性と Buck [5] の意味での可測性は精確に一致する．よって正則可測性は測密度の多次元 ($\#(A) \geq 2$) かつ非可換 (一般に $u, v \in A^*$ で $uv \neq vu$) な一般化と見なすことができる．しかし，[21] では概念の一般化だけが動機ではなく，原始語予想と呼ばれる具体的かつ組合せ論的な未解決問題へのアプローチとして正則可測性を導入し，いくつか非自明な結果を得ている (次節の補足 17 と補足 21 で簡単に説明する)．

[21] で導入した正則可測性は萌芽的な概念であり，まだまだ不明な点が多い．特に

言語が正則可測であるとはどういうことか？

については，ほとんど明らかになってないと言って良い．[21] では様々な正則可測・正則非可測な具体例を構成しているが，一般の言語クラス $\mathcal{C}$ に対する $\mathcal{C}$ 可測な言語の性質 (例えば $\mathcal{C}$ 可測な言語全体のなすクラスの閉包性や決定可能性，可測性の別の特徴づけ，などなど) については考察されていなかった．本論文ではこれら基礎的な性質について注目し議論を進めていく．

貢献

本研究は [21] の継続であり，2.3 節では [21] の結果も簡単に解説する．本論文において引用のない定理・系は（著者の知る限り）全て新しい定理であり，本論文での貢献はおもに以下の 4 つからなる：

1. 密度を持たない言語・正則可測な言語の具体例 (定理 7，定理 18–19)
2. $\mathcal{C}$ 可測な言語全体の閉包性，正則可測性の (条件付きの) 決定不能性および $\mathcal{C}$ 可測性の Carathéodory 条件による特徴づけ (定理 22–26).
3. 正則言語のいくつかの部分クラス $\mathcal{C}$ に対する $\mathcal{C}$ 可測性の考察 (定理 39–系 47)
4. 密度や $\mathcal{C}$ 可測性に関する問題と研究方針の提案 (予想 24，問題 49–51).

構成

本論文の構成は本節を除くと 4 つの節からなり，それぞれ上記の貢献 (1)–(4) に対応する：続く 2 節では言語の密度および可測性を定義し，密度を持たない言語・正則可測な言語の具体例を紹介する．その後 3 節で一般の言語クラス $\mathcal{C}$ に対して $\mathcal{C}$ 可測な言語全体のなすクラスの閉包性や Carathéodory 条件による特徴づけについて議論する．また、文脈自由言語に対する正則可測性の (ある予想の上での) 決定不能性を示す．4 節では，局所多様体と呼ばれる良い閉包性を満たす正則言語のいくつかの部分クラス $\mathcal{C}$ について注目し，それらのクラスにおける $\mathcal{C}$ 可測性を明らかにしていく．4 節では代数的言語理論の道具が必要になるため，簡単な入門 (4.1 節) を含めた．本研究のまとめは 5 節で行い，それまでに得られた定理や予想から今後の課題や方針を述べて終わる．

本論文全体において，形式言語理論や代数的言語理論の知識がなくてもなるべく話の流れがわかるよう，基礎的な定義・概念は可能な限り省略せずに書き例も多めに載せている．

2　形式言語の密度と可測性

形式言語理論の必要最低限な基本用語を 2.1 節で導入する．続く 2.2 節で言語の密度を導入し，最後に 2.3 節で可測性を定義し正則可測・正則非可測な言語の例をいくつか紹介する．

2.1 形式言語の基本用語

形式言語理論において，アルファベットとは単に空でない有限集合のことを差し，アルファベットの要素を文字と呼ぶ．以降，変数 A は常にアルファベットを表す．A 上の語とは A に属する文字 a_i を有限個並べた列：$a_1a_2\cdots a_n$ である．語には連接 $\cdot$ という演算を考えることができ，2 つの語 $u = a_1\cdots a_i, v = b_1\cdots b_j$ の連接 $u\cdot v$ は u と v を並べたものを表す：$u\cdot v = a_1\cdots a_ib_1\cdots b_j$. 語 w を n 個連接した文字列を w^n で表す．例えば $a^3 = aaa$ で $(ab)^2 = abab$ である．語 $w = a_1\cdots a_n$ の長さを $|w|$ で表す：$|w| = n$．ε で長さが 0 の語 (空語) を表す．集合 X の濃度を $\#(X)$ で表し，X が無限集合の場合は $\#(X) = \infty$ とする (本論文に現れる無限集合は全て可算である)．また集合 X の補集合を $\overline{X}$ で表す．語 $w = a_1\cdots a_n$ 中の $a \in A$ の個数を $|w|_a$ で表す：$|w|_a = \#\{i \mid a_i = a\}$. A 上の全ての語の集合を A^* で表し，A 上の長さが n である語の全ての集合を A^n で表す．語 w と u について，ある $x, y \in A^*$ が存在して $w = xuy$ が成り立つときに w は u を部分語に含むと言う．言語 L に対して $L \cap A^*wA^* = \emptyset$，すなわち L に属するどの語も w を部分語に含まないとき，L は w を禁句に持つと言う．逆に L が任意の語 w に対して $L \cap A^*wA^* \neq \emptyset$ となるとき，L は**稠密** (*dense*)[2] であると言う．

A 上の言語とは 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．任意の 2 つの言語 L, M の連接 $L\cdot M$ は以下のように語の連接を拡張して得られる：$L\cdot M = \{uv \in A^* \mid u \in L, v \in M\}$. 言語 L に対し，L^n で言語 L の n 回の連接 $L^0 = \{\varepsilon\}, L^n = L\cdot L^{n-1}$ を表す．言語 L に対して，その Kleene 閉包 L^* は $L^* = \cup_{n=0}^{\infty} L^n$ で定義される．A 上の言語 L に対し，文字列 $u \in A^*$ による左商 $u^{-1}L$ と右商 Lu^{-1} は次のように定義される：

$$u^{-1}L = \{v \in A^* \mid uv \in L\} \qquad Lu^{-1} = \{v \in A^* \mid vu \in L\}$$

本論文における言語の**クラス** $\mathcal{C}$ とは全てのアルファベットで添字付けられた言語の族 $\mathcal{C} = \{\mathcal{C}_A\}_{A:\text{アルファベット}}$ で各アルファベット $A \subseteq B$ に対して $\mathcal{C}_A$ の要素は A 上の言語 ($\mathcal{C}_A \subseteq 2^{A^*}$) かつ $\mathcal{C}_A \subseteq \mathcal{C}_B$ を満たすものである．ある A について $L \in \mathcal{C}_A$ であることを単に $L \in \mathcal{C}$ と書けることにする．以下に，本論文の考察の中心となる言語クラスを導入する．

定義 1 (正則言語，有限・補有限言語，星無し言語)**.** 言語 L について濃度が有限 ($\#(L) < \infty$) の場合は L を有限な言語と呼ぶ．言語 L が有限な言語に対する有限回の和集合・連接・Kleene 閉包の適用で表現できるとき，L を**正則言語** (*regular language*) と呼ぶ．正則言語全体のクラスを REG で表す．ぶ．L の補集合 $\overline{L}$ が有限 ($\#\left(\overline{L}\right) < \infty$) の場合は L を補有限な言語 (*cofinite language*) と呼ぶ．有限または補有限な言語全体のクラスを FIN で表す．言語 L が有限な言語に対する有限回の Bool 演算 (和集合・積集合・補集合) および連接の適用で表現できるとき，L を**星無し言語** (ほしなし言語，*star-free language*) と呼ぶ．星無し言語全体のクラスを SF で表す．

正則言語のクラス REG は補集合に閉じているため，星無し言語は定義より正則言語でもある．REG や FIN についてはよく知られているため例は不要であろうが，星無し言語 SF の例をいくつか挙げる．

例 2. 星無し言語はその名の通り Kleene 閉包 (演算 * は “Kleene star” とも呼ばれる) を使わず，その代わりに補集合が使える体系で表現できる正則言語である．空集合 $\emptyset$ は有限集合であるため星無し言語であり，その補集合 $A^* = \overline{\emptyset}$ も星無し言語である．任意の文字列 w に対して単元集合 $\{w\}$ は有限集合であるため星無し言語である．$\{w\}$ と A^* はそれぞれ星無し言語であるため，それらの連接 wA^*, A^*w, A^*wA^* すなわち「w で始まる語全体」「w で終わる語全体」「w を部分語

[2] 位相空間論において位相空間 X の部分集合 S が X において稠密であるとは，S の閉包 (S を含む最小の閉集合) が X と一致することを言う．これは言語の場合において直感的には「L の部分語全体の集合 $\{v \mid$ ある u, w が存在して $uvw \in L\}$ が A^* となる」すなわち稠密であることに対応している．より技術的に，A^* を空間とみなし空集合と任意の語 w について A^*wA^* を開基として考えることで A^* に位相が入り，言語 L が禁句を持たないという意味で稠密であることと，位相空間の意味で稠密であることが一致する (*cf.* [3] の演習問題 5.3).

に含む語全体」もそれぞれ星無し言語である．そのため，アルファベット $A = \{a, b\}$ に対して $(ab)^* = \{\varepsilon, ab, abab, ababab, \ldots\}$ は一見有限言語 $\{ab\}$ に Kleene 閉包を使っているため星無し言語ではないように思えるが，実は

$$(ab)^* = \{\varepsilon\} \cup (\overline{A^*aaA^*} \cap \overline{A^*bbA^*} \cap aA^* \cap A^*b)$$

と有限言語 $\{\varepsilon\}$ と星無し言語 $A^*aaA^*, A^*bbA^*, aA^*, A^*b$ に対する有限回の補集合・和集合・積集合の適用で表現できるため，$(ab)^*$ は星無し言語である．その一方，$(aa)^*$ は星無し言語では**ない**が，その証明は難しく，4 節で導入する Schützenberger の定理 (定理 32) を用いて示すことができる．特に，$(ab)^*$ や aA^* は有限でも補有限でもなく，$(aa)^*$ は明らかに正則言語であるため，FIN $\subsetneq$ SF $\subsetneq$ REG が成り立つ．

2.2 密度

定義 3 (自然密度と密度 (*cf.* [21]))**.** A 上の言語 L の**自然密度** (*natural density*) とは次の極限

$$\delta_A(L) = \lim_{n\to\infty} \frac{\#(L \cap A^n)}{\#(A^n)}$$

であり，極限が収束しない場合は「L は自然密度を持たない」と言い $\delta_A(L) = \perp$ で表す．A 上の言語 L の**密度** (*density*) とは次の極限

$$\delta^*_A(L) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} \frac{\#\big(L \cap A^i\big)}{\#(A^i)}$$

であり，極限が収束しない場合は「L は密度を持たない」と言い $\delta^*_A(L) = \perp$ で表す．

補題 4 ([21])**.** 密度を持つ言語 $K, L \subseteq A^*$ について以下が成り立つ．

(単調性) $K \subseteq L$ なら $\delta^*_A(K) \leq \delta^*_A(L)$.

(差集合) $K \subseteq L$ なら $\delta^*_A(L \setminus K) = \delta^*_A(L) - \delta^*_A(K)$．特に，$\delta^*_A(\overline{K}) = 1 - \delta^*_A(K)$.

(劣加法性) $\delta^*_A(K \cup L) \neq \perp$ なら $\delta^*_A(K \cup L) \leq \delta^*_A(K) + \delta^*_A(L)$.

(加法性) $K \cap L = \emptyset$ なら $\delta^*_A(K \cup L) = \delta^*_A(K) + \delta^*_A(L)$.

(連接) 任意の $w \in A^+$ に対して $\delta^*_A(wL) = \delta^*_A(L)/\#(A)^{|w|}$

初歩的な解析により，L が自然密度を持つ場合は密度も持ち，さらにその値は一致することがわかる．しかし，一般に逆は成り立たない．言語の密度の例をいくつか見てみよう．

例 5. 例えば言語 $L = (AA)^*$ に対しては $\delta_A(L)$ は収束しない (自然密度を持たない) が，$\delta^*_A(L)$ は $1/2$ に収束する (密度を持つ)．言語 L が有限ならば定義より明らかにその自然密度は 0 になる．逆に，L が補有限ならば自然密度は 1 となる．

任意の語 w に対して，A^*wA^* すなわち「w を部分語に含む語全体の集合」の密度は 1 となる．この事実は**無限の猿定理**と呼ばれている．L が禁句を持つ場合，すなわちある $w \in A^*$ で $L \cap A^*wA^* = \emptyset$ が成り立つときには，$A^*wA^* \subseteq \overline{L}$ のため無限の猿定理より $\delta^*_A(\overline{L}) = 1$ よって $\delta^*_A(L) = 0$ が成り立つ．その一方，禁句を持たないなら密度が正になるかと言えばそういうことはない．例えば半 Dyck 言語 $\mathsf{D} = \{w \in \{a, b\}^* \mid |w|_a = |w|_b$かつ任意の $uv = w$ について $|u|_a \geq |u|_b\}$[3]の密度は 0 である (簡単な考察から $\#\big(\mathsf{D} \cap A^{2n}\big)$ が n 番目の Catalan 数になることがわかり，n 番目の Catalan 数のオーダーが $\Theta(4^n/n^{3/2})$ となる事実からわかる) が，任意の語 $w \in \{a, b\}^*$ についてある $n, m \in \mathbb{N}$ が取れて $a^nwb^m \in \mathsf{D}$ とすることができるため D は禁句を持たない．

[3] a を開き括弧「(」だと思い，b を閉じ括弧「)」だと考えると，$w \in A^*$ が D に属すというのは「w はきちんと括弧の対応が取れている」ことに他ならない

注意 6. 密度はアルファベットに依存する：一般に，A 上で密度を持たない言語 L ($\delta^*_A(L) = \perp$) であっても，$B = A \cup \{c\}$ ($c \notin A$) において $L \cap B^*cB^* = \emptyset$ となり無限の猿定理から $\delta^*_B(L) = 0$ と B 上では密度を持つ．以降，単に「L が密度を持つ」と言った場合には，常に $L \subseteq A^*$ となる最小のアルファベット A 上で $\delta^*_A(L) \neq \perp$ となることを意味する．

次は密度を持たない言語の例である．

定理 7. 任意のアルファベット A について，言語

$$L_\perp = \{w \in A^* \mid \text{ある偶数 } n \text{ に対して } 3^n \leq |w| < 3^{n+1}\}$$

は密度を持たない．

Proof. $L_\perp$ の密度は定義より

$$\frac{1}{n}\sum_{i=0}^{n-1} \frac{\#\big(L_\perp \cap A^i\big)}{\#(A^i)} \tag{1}$$

の $n \to \infty$ での極限である．定義より，ある偶数 k に対して長さが 3^k 以上 3^{k+1} 未満の語は全て $L_\perp$ に属するため，$n = 3^k$ の分数 (1) の値を $0 \leq \alpha \leq 1$ とすると $n = 3^{k+1}$ での分数 (1) の値は

$$\frac{1}{3^{k+1}}\left(\sum_{i=0}^{3^k-1} \frac{\#\big(L_\perp \cap A^i\big)}{\#(A^i)} + \sum_{i=3^k}^{3^{k+1}-1} 1\right) = \frac{1}{3^{k+1}}(3^k\alpha + 3^{k+1} - 3^k) = \frac{\alpha + 3 - 1}{3} = \frac{\alpha + 2}{3} \geq 2/3 \tag{2}$$

となる．逆に，ある奇数 k に対して長さが 3^k 以上 3^{k+1} 未満の語は全て $L_\perp$ に属さないため，$n = 3^k$ の分数 (1) の値を $0 \leq \beta \leq 1$ とすると $n = 3^{k+1}$ での分数 (1) の値は

$$\frac{1}{3^{k+1}}\left(\sum_{i=0}^{3^k-1} \frac{\#\big(L_\perp \cap A^i\big)}{\#(A^i)} + \sum_{i=3^k}^{3^{k+1}-1} 0\right) = \frac{3^k\beta}{3^{k+1}} = \frac{\beta}{3} \leq 1/3 \tag{3}$$

となり，値 (1) は 2/3 以上の値と 1/3 以下の値をそれぞれ無限回取るため収束せず，したがって $L_\perp$ は密度を持たない． □

このように，言語によっては密度を持たないものもあるが，次の定理は任意の正則言語は密度を持つことを保証する．

定理 8 (*cf.* [18] の Theorem III.6.1)**.** L を A 上の正則言語とする．このときある自然数 c が存在して，$d < c$ となる各自然数 d に対して次の極限が収束して有理数となる：

$$\lim_{n\to\infty} \frac{\#\big(L \cap A^{cn+d}\big)}{\#(A^{cn+d})}$$

系 9. 任意の正則言語は密度を持つ．

系 10. 任意の正則言語 $L \subseteq A^*$ に対して $\delta_A(L) = 0$ と $\delta^*_A(L) = 0$ は等価である．

さらに，正則言語においては「密度が 0 である」という「疎」概念は,「禁句を持つ」という位相的な「疎」概念と一致するという理論的に良い性質がある．

定理 11 ([20])**.** 正則言語 L について，$\delta^*_A(L) = 0$ であることと，L は禁句を持つことは同値.

2.3 可測性

定義 12 (内密度，外密度，可測性)**.** $\mathcal{C}, \mathcal{D}$ を言語クラスとする．A 上の言語 L に対する (A 上の)$\mathcal{C}$ **内密度**を

$$\underline{\mu}_{\mathcal{C}_A}(L) = \sup\{\delta^*_A(K) \mid K \subseteq L, K \in \mathcal{C}_A, \delta^*_A(K) \neq \bot\}$$

で定義し，(A 上の)$\mathcal{C}$ **外密度**を

$$\overline{\mu}_{\mathcal{C}_A}(L) = \inf\{\delta^*_A(K) \mid L \subseteq K, K \in \mathcal{C}_A, \delta^*_A(K) \neq \bot\}$$

で定義する．L が (A 上)$\mathcal{C}$ **可測**であるとは，$\overline{\mu}_{\mathcal{C}_A}(L) = \underline{\mu}_{\mathcal{C}_A}(L)$ が成り立つことを言い，このとき $\overline{\mu}_{\mathcal{C}_A}(L)$ を $\mu_{\mathcal{C}_A}(L)$ で表し L の (A 上の)$\mathcal{C}$ **測度**と呼ぶ．$\mathcal{D}$ が $\mathcal{C}$ 可測であるとは，各アルファベット A に対して $\mathcal{D}_A$ に属する全ての言語が A 上 $\mathcal{C}$ 可測であることを言う．

注意 13. $\mathcal{C}_A$ が空集合と A^* を含む場合は上記定義の inf および sup の右辺の集合は常に非空であり $0 \leq \underline{\mu}_{\mathcal{C}_A}(L) \leq \overline{\mu}_{\mathcal{C}_A}(L) \leq 1$ が成り立つ．密度と同様に，可測性はアルファベットに依存する (例 33 でも改めて説明する)．以降，単に「L が $\mathcal{C}$ 可測である」と言った場合には，常に $L \subseteq A^*$ となる **最小の**アルファベット A 上で可測であることを意味することとし，$\overline{\mu}_{\mathcal{C}_A}(L), \underline{\mu}_{\mathcal{C}_A}(L), \mu_{\mathcal{C}_A}(L)$ を単に $\overline{\mu}_{\mathcal{C}}(L), \underline{\mu}_{\mathcal{C}}(L), \mu_{\mathcal{C}}(L)$ と A を省略して書けることとする．

さらに，以降，本論文では常に全ての言語が密度を持つようなクラスのみについて考える．そのため，単に「言語クラス $\mathcal{C}$」と言った場合は,「任意の $L \in \mathcal{C}$ について L は密度を持つ」と暗に仮定する．実際，以降は具体的な $\mathcal{C}$ について考察する際には常に $\mathcal{C}$ は正則言語のクラス REG かその部分クラスのみを考える．また,「REG 可測」を「正則可測」と呼ぶこととする．

続く 2.4 節で正則可測・正則非可測なさまざまな言語の例を紹介するが，ここでは一般に「言語 L が $\mathcal{C}$ 可測である」ことの直感を説明する．言語 L に対してある $K \in \mathcal{C}$ が存在して $K \subseteq L$ かつ $\delta^*_A(L) - \delta^*_A(K) = 0$ となる場合に「L は $\mathcal{C}$ で**内側から零近似可能**」と呼ぶ．同様に，$\delta^*_A(M) - \delta^*_A(L) = 0$ なる $L \subseteq M \in \mathcal{C}$ が存在するときに「L は $\mathcal{C}$ で**外側から零近似可能**」と呼ぶ．$K \subseteq L \subseteq M$ であれば定義より $\delta^*_A(K) \leq \underline{\mu}_{\mathcal{C}}(L) \leq \overline{\mu}_{\mathcal{C}}(L) \leq \delta^*_A(M)$ であるから，上のように L が $\mathcal{C}$ で内外から (K と M で) 零近似可能である場合は $\delta^*_A(K) = \underline{\mu}_{\mathcal{C}}(L) = \overline{\mu}_{\mathcal{C}}(L) = \delta^*_A(M)$ となり L は $\mathcal{C}$ 可測となる．

しかし，L が内外から $\mathcal{C}$ で零近似可能で**ない**場合でも $\mathcal{C}$ 可測になる場合があり (続く 2.4 でそのような例を説明する)，$\mathcal{C}$ 可測性の議論ではどちらかというとこちらの場合が重要になるため，用語を導入する．言語 L と言語の列 $(K_n)_n$ について (1) 各 $n \in \mathbb{N}$ で $K_n \subseteq L$ (2) $\lim_{n\to\infty} \delta^*_A(L) - \delta^*_A(K_n) = 0$ となる場合「**列 $(K_n)_n$ は内側から L に収束する**」と言う．条件 (1) の包含関係を逆 (各 $n \in \mathbb{N}$ で $L \subseteq K_n$) にすることで「**列 $(K_n)_n$ は外側から M に収束する**」も同様に定義できる．$\mathcal{C}$ の言語の列 $(K_n)_n$ で L に内側から収束するような列が存在する場合は，定義より $\underline{\mu}_{\mathcal{C}}(L) = \lim_{n\to\infty} \delta^*_A(K_n) = \delta^*_A(L)$ が成り立つ．同様に，$\mathcal{C}$ の言語の列 $(M_n)_n$ で L に外側から収束するような列が存在する場合は，定義より $\overline{\mu}_{\mathcal{C}}(L) = \lim_{n\to\infty} \delta^*_A(M_n) = \delta^*_A(L)$ が成り立つ．そのため L に内側・外側から収束する $\mathcal{C}$ の言語の列 $(K_n)_n, (M_n)_n$ が存在する場合は $\underline{\mu}_{\mathcal{C}}(L) = \overline{\mu}_{\mathcal{C}}(L)$ となり L は $\mathcal{C}$ 可測となる．このとき，$\mathcal{C}$ の言語の対の列 $(K_n, L_n)_n$ は L に**両側から収束する**と言う．実際，L が $\mathcal{C}$ 可測であることと，L に両側から収束する $\mathcal{C}$ の言語の対の列が存在することは等価である．内外測度の基本的な性質を述べた以下の補題は次節で用いる．

補題 14 ([21])**.** $\mathcal{C}$ を補集合に閉じた言語クラスとする．このとき言語 L が $\mathcal{C}$ 可測であることと，$\overline{\mu}_{\mathcal{C}}(L) + \overline{\mu}_{\mathcal{C}}(\overline{L}) = 1$ が成り立つことは同値である．

補題 15 ([21])**.** クラス $\mathcal{C}$ と言語 K, L について次が成り立つ：

(不等式) $\delta^*_A(L) \neq \bot$ ならば $\underline{\mu}_{\mathcal{C}}(L) \leq \delta^*_A(L) \leq \overline{\mu}_{\mathcal{C}}(L)$．$L$ が $\mathcal{C}$ 可測ならば $\underline{\mu}_{\mathcal{C}}(L) = \delta^*_A(L) = \overline{\mu}_{\mathcal{C}}(L)$.

(単調性) $K \subseteq L$ ならば $\overline{\mu}_{\mathcal{C}}(K) \leq \overline{\mu}_{\mathcal{C}}(L)$.

(劣加法性) $\mathcal{C}$ が和集合について閉じてるならば $\overline{\mu}_{\mathcal{C}}(K \cup L) \leq \overline{\mu}_{\mathcal{C}}(K) + \overline{\mu}_{\mathcal{C}}(L)$.

$\overline{\mu}_{\mathcal{C}}(L) - \underline{\mu}_{\mathcal{C}}(L)$ の値は 0 から 1 の間の実数値であり，直感的には言語 L の「$\mathcal{C}$ での近似しにくさ」を表していると考えて良い．この値 $\overline{\mu}_{\mathcal{C}}(L) - \underline{\mu}_{\mathcal{C}}(L)$ を L **の** $\mathcal{C}$ **ギャップ**と呼ぶ．L の $\mathcal{C}$ ギャップが 0 であるとは L が可測ということであり，逆に $\mathcal{C}$ ギャップが 1 となる場合は L が $\mathcal{C}$ によって内側からも外側からも密度の意味で「全く近似できない」という状況を表す.

2.4　正則可測・正則非可測な言語の例

定理 16 ([21])**.** 以下の文脈自由言語は，全て正則言語ではないが，全て正則可測：

1. $\mathsf{D} = \{w \in \{a,b\}^* \mid |w|_a = |w|_b$かつ任意の $uv = w$ について $|u|_a \geq |u|_b\}$.
2. $\mathsf{B} = \{w \in \{a,b\}^* \mid |w|_a = |w|_b\}$.
3. $\mathsf{P} = \{w \in \{a,b\}^* \mid w$ は左から読んでも右から読んでも同じ語 (すなわち回文)$\}$.
4. $\mathsf{O}_3 = \{w \in \{a,b,c\}^* \mid |w|_a = |w|_b$または $|w|_a = |w|_c\}$.
5. $\mathsf{O}_4 = \{w \in \{x,\overline{x},y,\overline{y}\}^* \mid |w|_x = |w|_{\overline{x}}$または $|w|_y = |w|_{\overline{y}}\}$.
6. $\mathsf{G} = \{a^{n_1}ba^{n_2}b\cdots a^{n_k}b \mid k \geq 1$ で各 $1 \leq i \leq k$ に対して $n_i \neq i\}$.
7. $\mathsf{K} = S_1\{c\}A^* \cup S_2\{c\}A^*$，ここで $A = \{a,b,c\}$ で S_1 と S_2 はそれぞれ
$$S_1 = \{a\}\{b^ia^i \mid i \geq 1\}^* \qquad S_2 = \{a^ib^{2i} \mid i \geq 1\}^*\{a\}^+$$
と定義されている.

補足 17. 定理 16 中に表れる言語は全て文脈自由言語であるが，いずれも正則言語ではない．$\mathsf{D}, \mathsf{B}, \mathsf{P}$ は典型的な非正則言語の例としてよく知られているが，それ以外の言語についてその複雑性を簡単に説明する.

O_3 と O_4 は Flajolet [12] によって**本質的に曖昧**[4]な文脈自由言語であることが示されており，文脈自由言語の中でも比較的複雑な構造を持つ.

古典的な Chomsky–Schützenberger の定理 [6] は，もし文脈自由言語 L が無曖昧ならば，次で定義される級数
$$F_L(z) = \sum_{n=0}^{\infty} \#(L \cap A^n)\, z^n$$
すなわち L の**母関数**が代数関数となることを述べている．よって L の母関数が超越関数となるならば，L は本質的に曖昧となるが，Flajolet は O_4 の母関数が超越関数である [12] こと，G の母関数が超越関数である [13] ことをそれぞれ示している.

最後に，Kemp [15] によって定義された言語 K の複雑さについて説明する．言語の密度はもともと代数的符号理論 (*cf.* [4]) などの文脈で Berstel らによって考察がなされていたが，Berstel は彼の 1973 年の論文 [2] にて任意の無曖昧文脈自由言語の密度が代数的数となることを示した．[2] 以降，密度が超越数であるような文脈自由言語が存在するかどうかは未解決であったが，1980 年に Kemp [15] がそのような言語を実際に構成した．それが定理 16 の K である.

このように，G や K は形式言語理論的な視点 (本質的に曖昧) からも組合せ論的な視点 (母関数・密度がそれぞれ超越的) からも複雑な文脈自由言語であり，それにも関わらず正則言語でいくらでも近似ができるということを定理 16 は言っている.

[4] 文脈自由言語 L が無曖昧 (*unambiguous*) であるとは，L を生成する無曖昧な文脈自由文法が存在することを言う．文脈自由言語が本質的に曖昧とは，それを生成する無曖昧な文法が存在しないことを言う．文法が無曖昧であるとは，任意の語 w についてその文法において w を生成する最左導出列がたかだか 1 つであることを言う．無曖昧な文脈自由言語の詳細は文献 [23] の 3 節を参照されよ.

Proof. 定理 16 の証明は全て [21] に書かれているが，可測性の良い例題となるため，半 Dyck 言語 D が正則可測であることの証明の概略をここに記す．$A = \{a, b\}$ とする．まず，例 5 で述べたように D の密度は 0 であるため，$\emptyset$ で内側から自明に零近似できる．しかし，D は正則言語で外側から零近似できない．なぜなら，定理 11 より正則言語 L の密度が 0 であれば禁句を含むためある $w \in A^*$ で $L \cap A^* w A^* = \emptyset$ となる．しかし，例 5 で説明したように D は禁句を持たない (稠密である)．よって $\mathsf{D} \cap A^* w A^* \neq \emptyset$ となり $\mathsf{D} \subsetneq L$ であるため L は D の外側からの零近似にはなりえない．

よって D が正則可測であることを示すためには，D に外側から収束する正則言語の列を構成する必要がある．各 $k \geq 1$ に対して $L_k = \{w \in \{a, b\}^* \mid |w|_a = |w|_b \mod k\}$ と定義する．L_k は k 状態の決定性オートマトンで認識できるため，正則言語であり，特に $\mathsf{D} \subseteq L_k$ が各 k について成り立つ．簡単な解析から L_k の密度が $1/k$ となることが示せ，よって正則言語の列 $(L_k)_{k \geq 1}$ は D に外側から収束する． □

また，次の言語 L_{2^n} は文脈依存言語ではあるが，文脈自由言語ではない．この言語も正則可測である．

定理 18. 言語 $L_{2^n} = \{a^{2^n} \in \{a\}^* \mid n \in \mathbb{N}\}$ は正則可測である．

Proof. $A = \{a\}$ とする．$\delta_A^*(L_{2^n}) = 0$ が成り立つため，外側から収束する正則言語の列を構成すれば良い．$L_k = (a^k)^* \cup \{a^n \mid 0 < n < k\}$ は正則言語であり，各 $k \geq 1$ に対して $\delta_A^*(L_k) = 1/k$ が成り立ち，特に $\lim_{k \to \infty} \delta_A^*(L_k) = 0$ である．任意の $k \geq 1$ と $n \in \mathbb{N}$ について，$a^{2^n} \in L_{2^k}$ すなわち $L_{2^n} \subseteq L_{2^k}$ であることを示す．$2^n < 2^k$ ならば定義より明らかに L_{2^k} に属するので，$2^n \geq 2^k$ の場合を考えればよいが，$2^n = 2^k \cdot 2^{n-k}$ であるため a^{2^n} は a^{2^k} の $2^{n-k} \in \mathbb{N}$ 回の繰り返しとなり $(a^{2^k})^*$ に属する．よって正則言語の列 $(L_{2^k})_{k \geq 1}$ は L_{2^n} に外側から収束する． □

さらに，任意の実数 $0 \leq \alpha \leq 1$ に対して，$\delta_A^*(L) = \alpha$ となる正則可測な言語 L が存在することが示せる．

定理 19. A が 2 つ以上の文字を含む場合，任意の実数 $0 \leq \alpha \leq 1$ に対して $\delta_A^*(L) = \alpha$ となる正則可測な言語 L が存在する．

Proof. $A = \{a, b\}$ の場合について考える (一般の場合も同様に示せる)．実数 $\alpha \in [0, 1]$ の 2 進数展開を考えると

$$\alpha = \sum_{n=1}^{\infty} \alpha_n 2^{-n}$$

となる列 $(\alpha_n)_{n \geq 1}$(ただし各 i で $\alpha_i \in \{0, 1\}$) が存在する．$K_0 = \emptyset, M_0 = A^*$ として，各 $n \geq 1$ について正則言語 K_n, M_n を以下のように定義する：

$$K_n = \begin{cases} b^{n-1} a A^* \cup K_{n-1} & \alpha_n = 1 \\ K_{n-1} & \alpha_n = 0 \end{cases} \qquad M_n = \begin{cases} M_{n-1} & \alpha_n = 1 \\ M_{n-1} \setminus b^{n-1} a A^* & \alpha_n = 0 \end{cases}$$

各 $n, m \geq 1$ に対して $n \neq m$ ならば $b^{n-1} a A^* \cap b^{m-1} a A^* = \emptyset$ のため，構成から任意の $n \in \mathbb{N}$ で $K_n \subseteq M_n$ が成り立つ．密度の連接の性質 (補題 4) から各 $n \geq 1$ に対して $\delta_A^*(b^{n-1} a A^*) = 2^{-n}$ が成り立つ．また，密度の加法性および差集合の性質 (補題 4) より各 $n \geq 1$ に対してそれぞれ

$$\delta_A^*(K_n) = \sum_{i=1}^{n} \alpha_n 2^{-n} \qquad \delta_A^*(M_n) = 1 - \sum_{i=1}^{n} (1 - \alpha_n) 2^{-n}$$

が成り立つ．よって $\lim_{n \to \infty} \delta_A^*(K_n) = \delta_A^*(M_n) = \alpha$ となり，$(K_n, M_n)_n$ は密度が $\delta_A^*(L) = \alpha$ となる言語 $L = \bigcup_{n \in \mathbb{N}} K_n = \bigcap_{n \in \mathbb{N}} M_n$ に両側から収束する． □

このように，多くの複雑な言語が正則可測であり，さらに正則可測な言語は非可算無限個存在する一方，次の言語は正則**非**可測である．

定理 20 ([21])**.** $A = \{a, b\}$ 上の以下の言語は全て正則非可測：

1. 各 $n \geq 1$ について $\mathsf{M}_n = \{w \in A^* \mid |w|_a > n \cdot |w|_b\}$.
2. 原始語全体の集合 Q．ここで w が**原始語** (*primitive word*) であるとは，w が非空であり，自分よりも短い語の繰り返しとならない ($v^n = w \Rightarrow v = w$ かつ $n = 1$) ことを言う．

特に，Q と $n \geq 2$ の場合の M_n はそれぞれ正則ギャップが 1 となり，強い意味で正則非可測である．

補足 21. M_1 は「a の個数が b の個数よりも多い $\{a, b\}$ 上の文字列全体」であり，これが正則言語によって「近似できない」ことは最初に Eismann–Ravikumar [11] によって証明された．著者は [21] において Eismann–Ravikumar の証明を単純化し，さらに $n \geq 2$ の場合においても M_n が正則非可測であることを示した．各 $n \geq 1$ について M_n は決定性プッシュダウンオートマトンで受理できるため，決定的な文脈自由言語である．

原始語の集合 Q は語の組合せ論や文字列アルゴリズムにおいて非常に重要な対象であるが「Q は文脈自由ではない」という Dömösi–Horváth–Ito 予想 [9] は 30 年来の未解決問題である (この予想に関するこれまでの既存研究は書籍 [7] で網羅的にまとめられている)．著者はこの予想を肯定的に解決するために「任意の文脈自由言語の正則ギャップは 1 未満である一方，Q の正則ギャップは 1 になるのではないか」と言う予想を立てた．Q の正則ギャップが 1 になることは示せたが，その後 M_n $(n \geq 2)$ という反例 (文脈自由言語ながら正則ギャップが 1) を見つけたというのが [21] の研究背景である．「Q は CFL 非可測である ([21])」(よって Q は文脈自由ではない) と著者は予想しているが，こちらはまだ証明も反証もされていない．

3　可測な言語族の一般的な性質

前節では正則可測・非可測な言語の具体例をいくつか見てきた．本節では可測な言語族の一般的な性質，すなわち可測な言語全体のクラスの閉包性，および正則可測性の決定不能性について議論する．最後に，Carathéodory 条件による可測性の特徴づけも与える．

定理 22. $\mathcal{C}$ を Bool 演算に閉じている言語クラスとする．L, K がそれぞれ $\mathcal{C}$ 可測であり，L, K と $\mathcal{C}$ の言語から有限回の Bool 演算の適用で構成される言語が密度を持つならば，補集合 $\overline{L}$，和集合 $L \cup K$ および積集合 $L \cap K$ もそれぞれ $\mathcal{C}$ 可測である．

Proof. $\mathcal{C}$ は補集合に閉じているため，補題 14 より L が $\mathcal{C}$ 可測であることと $\overline{L}$ が $\mathcal{C}$ 可測であることは同値である．よって L と K が $\mathcal{C}$ 可測のときに $L \cup K$ が $\mathcal{C}$ 可測であることを示せば十分である．L と K にそれぞれ内側から収束する $\mathcal{C}$ の列 $(L_n)_n, (K_n)_n$ をそれぞれ固定する．$\mathcal{C}$ は和集合に閉じているため $(L_n \cup K_n)_n$ は $\mathcal{C}$ の列となるがこれが内側から $L \cup K$ に収束することを示そう．任意の $\epsilon/2 > 0$ についてある δ が存在して $\delta < n$ なる n では $\delta^*_A(L) - \delta^*_A(L_n), \delta^*_A(K) - \delta^*_A(K_n) < \epsilon/2$ が成り立つため，特に密度の差集合の性質 (補題 4) から

$$(\delta^*_A(L) - \delta^*_A(L_n)) + (\delta^*_A(K) - \delta^*_A(K_n)) = \delta^*_A(L \setminus L_n) + \delta^*_A(K \setminus K_n) < \epsilon \tag{4}$$

が成り立つ．仮定より $(L \setminus L_n) \cup (K \setminus K_n)$ は密度を持つため，密度の劣加法性 (補題 4) から

$$\delta^*_A((L \setminus L_n) \cup (K \setminus K_n)) \leq \delta^*_A(L \setminus L_n) + \delta^*_A(K \setminus K_n) < \epsilon \tag{5}$$

が得られる．$(L \cup K) \setminus (L_n \cup K_n) \subseteq (L \setminus L_n) \cup (K \setminus K_n)$ のため $\delta^*_A((L \cup K) \setminus (L_n \cup K_n)) < \epsilon$ となり，すなわち列 $(L_n \cup K_n)$ が内側から $L \cup K$ に収束すること言っている．同様にして外側からの収束列も構成できるため $L \cup K$ は $\mathcal{C}$ 可測である． □

定理 23. $\mathcal{C}$ を左商 (または右商) に閉じている言語クラスとする．L が $\mathcal{C}$ 可測であり，L の左商 $a^{-1}L$(または右商 La^{-1}) が密度を持つのであれば $\mathcal{C}$ 可測である．

Proof. A 上の言語 L が $\mathcal{C}$ 可測として $a^{-1}L$ も可測であることを示す (右商も同様に示せる)．定義より L に両側から収束する $\mathcal{C}$ の言語の対列 $(K_n, M_n)_n$ が存在するが，このとき $(a^{-1}K_n, a^{-1}M_n)_n$ が $a^{-1}L$ に両側から収束することを示す．

簡単のために $A = \{a, b\}$ とする (一般の場合でも同様に示せる)．まず，各 $a \in A$ に対して $L \cap aA^* = aa^{-1}L$ が成り立つため，$L = aa^{-1}L \cup bb^{-1}L \cup (L \cap \{\varepsilon\})$ と書くことができる．仮定から $aa^{-1}L$ と $bb^{-1}L$ は密度を持ち，さらに共通部分を持たないため密度の加法性 (補題 4) から

$$\delta_A^*(L) = \delta_A^*(aa^{-1}L) + \delta_A^*(bb^{-1}L) \tag{6}$$

が成り立つ．定義より各 n に対して $K_n \subseteq L$ となるため，$a^{-1}K_n \subseteq a^{-1}L$ となり，特に $\delta_A^*(a^{-1}K_n) \leq \delta_A^*(a^{-1}L)$ であることがわかる．$(K_n)_n$ は L に内側から収束するため，任意の $\epsilon/2 > 0$ に対してある δ が存在して $\delta < n$ となる n では $\delta_A^*(L) - \delta_A^*(K_n) < \epsilon/2$ となるが，このとき，式 (6) より同時に

$$\delta_A^*(L) - \delta_A^*(K_n) = (\delta_A^*(aa^{-1}L) - \delta_A^*(aa^{-1}K_n)) + (\delta_A^*(bb^{-1}L) - \delta_A^*(bb^{-1}K_n)) < \frac{\epsilon}{2} \tag{7}$$

が成り立つ．密度の連接の性質 (補題 4) より，不等式 (7) は

$$\frac{1}{2}(\delta_A^*(a^{-1}L) - \delta_A^*(a^{-1}K_n)) + \frac{1}{2}(\delta_A^*(b^{-1}L) - \delta_A^*(b^{-1}K_n)) < \frac{\epsilon}{2}$$

と変形でき，ここから各 $n > \delta$ について $\delta_A^*(a^{-1}L) - \delta_A^*(a^{-1}K_n) < \epsilon$ が成り立つことが言え，すなわち $(a^{-1}K_n)_n$ は $a^{-1}L$ に内側から収束する列となっている．同様に，$(a^{-1}M_n)_n$ は $a^{-1}L$ に外側から収束する列となることも示せる． □

正則言語によって近似できる言語全体のクラスを

$$\mathrm{Ext(REG)} = \{L \subseteq A^* \mid L \text{ は } A \text{ 上正則可測 }\}_{A:\text{アルファベット}}$$

と書くことにする．定理 23 で示した閉包性および定理 20 から，文脈自由言語の正則可測性の決定不能性は（次の予想を仮定した上で）Greibach のメタ定理 [14][5]を用いることで示せる．なお文脈自由言語の定義は省略するが，必要であれば [23] の 3 節を参照されよ．

予想 24. 文脈自由言語 $L \subseteq A^*$ が密度を持つならば，その商 $a^{-1}L, La^{-1}$ も密度を持つ．

定理 25. 予想 24 が正しければ，与えられた文脈自由文法が正則可測な言語を生成するかどうかは決定不能．

Proof. 文脈自由言語全体のクラスを CFL で表す．CFL は左右からの商に付いて閉じている．また，仮定および定理 23 より正則可測な文脈自由言語全体 $P = \mathrm{Ext}_A(\mathrm{REG}) \cap \mathrm{CFL}$ もまた左右からの商について閉じている．明らかに $\mathrm{REG} \subseteq P$ が言え，また定理 20-(1) より正則非可測な文脈自由言語が存在するため $P \subsetneq \mathrm{CFL}$ である．CFL においては普遍性 (*universality*: 与えられた文脈自由文法について，その文法が A^* を生成するかどうか) が決定不能であるため，Greibach のメタ定理 [14] より，与えられた文法が正則可測な文脈自由言語を生成するかどうかも決定不能である． □

最後に，$\mathcal{C}$ 可測であることの Carathéodory 条件による特徴付けを与える．以下の定理の証明は，形式言語の密度特有の条件も入っているが，議論の骨子は通常の測度論における Lebesgue 可測性の Carathéodory 条件による特徴付けと同様である (*cf.* [22] の定理 5.1 の証明 (p.81))．

[5]紙面の都合上正確な言明は省略するが，Greibach のメタ定理は「(CFL のような) ある程度の複雑さを持つ言語族 $\mathcal{C}$ において，任意の正則言語で成立する性質 P がある $L \in \mathcal{C}$ において成立しない場合，与えられた $\mathcal{C}$ の言語 (の記述) が性質 P を満たすかどうかは決定不能」であることを言っている．

定理 26. Bool 演算に閉じたクラス $\mathcal{C}$ と言語 $L \subseteq A^*$ について考える．L および $\mathcal{C}$ の言語に対して有限回 Bool 演算を適用して得られる任意の言語が密度を持つ場合，L が $\mathcal{C}$ 可測であることと次の Carathéodory 条件は同値：

$$\forall X \subseteq A^* \qquad \overline{\mu}_{\mathcal{C}}(X) = \overline{\mu}_{\mathcal{C}}(X \cap L) + \overline{\mu}_{\mathcal{C}}(X \cap \overline{L}). \tag{8}$$

Proof. L が Carathéodory 条件 (8) を満たせば，特に $X = A^*$ と置くことで $\overline{\mu}_{\mathcal{C}}(A^*) = 1 = \overline{\mu}_{\mathcal{C}}(L) + \overline{\mu}_{\mathcal{C}}(\overline{L})$ となり，$\mathcal{C}$ は補集合に閉じているため補題 14 より L は $\mathcal{C}$ 可測であることが直ちに得られる．

L が $\mathcal{C}$ 可測であると仮定して Carathéodory 条件 (8) を満たすことを示す．任意の $X \subseteq A^*$ について，$\overline{\mu}_{\mathcal{C}}$ の定義より任意の $\epsilon > 0$ に対して $X \subseteq K$ かつ $\delta_A^*(K) \leq \overline{\mu}_{\mathcal{C}}(X) + \epsilon$ が成り立つ $K \in \mathcal{C}$ が存在する．L および $K, \overline{K} \in \mathcal{C}$ はそれぞれ $\mathcal{C}$ 可測であり，仮定より Bool 演算 $K \cap L$ と $K \cap \overline{L}$ はそれぞれ密度を持つため，定理 22 よりそれぞれ $\mathcal{C}$ 可測でもある．$K = (K \cap L) \cup (K \cap \overline{L})$ かつ $(K \cap L) \cap (K \cap \overline{L}) = \emptyset$ のため密度の加法性 (補題 4) より

$$\delta_A^*(K) = \delta_A^*(K \cap L) + \delta_A^*(K \cap \overline{L}) \tag{9}$$

が得られる．よって，

$$\overline{\mu}_{\mathcal{C}}(X) \geq \delta_A^*(K) - \epsilon = \delta_A^*(K \cap L) + \delta_A^*(K \cap \overline{L}) - \epsilon \tag{10}$$

$$\geq \overline{\mu}_{\mathcal{C}}(X \cap L) + \overline{\mu}_{\mathcal{C}}(X \cap \overline{L}) - \epsilon \tag{11}$$

となり，$\epsilon > 0$ は任意に取っていたため最終的に

$$\overline{\mu}_{\mathcal{C}}(X) \geq \overline{\mu}_{\mathcal{C}}(X \cap L) + \overline{\mu}_{\mathcal{C}}(X \cap \overline{L}) \tag{12}$$

が得られる．逆向きの不等式は $\overline{\mu}_{\mathcal{C}}$ の劣加法性 (補題 15) から直ちに言える． □

4 正則言語の局所多様体と可測性

本節では，局所多様体と呼ばれる良い閉包性を持った正則言語のいくつかの部分クラスについて，その可測性を明らかにしていく．正則言語の局所多様体の定義や，その有限モノイドとの対応 (Eilenberg 型の定理) の説明には代数的言語理論の基礎知識が不可欠のため，続く 4.1 節で簡単に入門する．

4.1 代数的言語理論の基礎と Eilenberg の多様体定理

代数的言語理論では，まず A^* を連接 $\cdot$ を二項演算として持った代数構造 (ε を単位元とした自由モノイド) とみなし各概念を定義していく．言語 $L \subseteq A^*$ について，その**統語的合同関係** (*syntactic congruence*) $\equiv_L$ は

$$v \equiv_L w \Leftrightarrow \forall x, y \in A^* (xvy \in L \text{ と } xwy \in L \text{ は同値})$$

と定義される A^* 上の二項関係であり，これは A^* 上の合同関係 (同値関係かつ $u \equiv_L w \Rightarrow \forall x, y \in A^*(xuy \equiv_L xwy)$) となる．$L \subseteq A^*$ の**統語モノイド** (*syntactic monoid*)$\mathsf{Synt}(L)$ とは自由モノイド A^* をこの合同関係 $\equiv_L$ で割ったモノイド $\mathsf{Synt}(L) = A^*/\equiv_L$ である．$\mathsf{Synt}(L)$ は A^* を割ったモノイドであるため，自然な全射準同型 $\eta_L : A^* \to \mathsf{Synt}(L)$ が導出されるが，これを L の**統語準同型** (*syntactic morphism*) と呼ぶ．定義より，任意の言語 L について $L = \eta_L^{-1}(\eta_L(L))$ が成り立つ．

例 27. $A = \{a\}$ として言語 $L = (aa)^*$ について考える．2 つの自然数 n, m の偶奇が等しい ($n = m \mod 2$) とき，任意の $x, y \in A^*$ について語 $xa^n y$ と語 $xa^m y$ の長さの偶奇は等しいため，それらが L に属するのは同値である．よって $a^n \equiv_L a^m$ が成り立つ．逆に，n, m の偶奇が異なるのであれば，a^n か a^m の偶数長の方だけが L に属するため $a^n \not\equiv_L a^m$ となる．よって L の統語モノイドは位数 2 の群と同型 $\mathsf{Synt}((aa)^*) \cong \mathbb{Z}/2\mathbb{Z}$ となる．

次に，$A=\{a,b\}$ として半 Dyck 言語 $\mathsf{D}\subseteq A^*$ について考える．語 w 中に表れる括弧の整合的な対応を意味する ab という部分語を全て消して得られる語を $\mathrm{Trim}(w)$ として表す．例えば $\mathrm{Trim}(a\underline{ab}b)=ab$ であり $\mathrm{Trim}(b\,\underline{ab}\,\underline{ab}\,a)=ba$ となる．簡単な考察から，任意の語 w についてその不動点 $\mathrm{Trim}^*(w)$ が一意に定まり $\mathrm{Trim}^*(w)=b^ma^n$ $(m,n\in\mathbb{N})$ という形となることが分かる (a のあとに b が来ればそれは Trim によって消されるので，不動点は a のあとに b は来ない)．$w\in\mathsf{D}$ と $\mathrm{Trim}^*(w)=\varepsilon$ は同値であり，実際，括弧の対応の整合性が取れているかどうかは $\mathrm{Trim}^*(w)$ のみが重要であり，特に $u\equiv_{\mathsf{D}}v\Leftrightarrow\mathrm{Trim}^*(u)=\mathrm{Trim}^*(v)$ が成り立つ．よって $\mathsf{Synt}(\mathsf{D})$ は 2 元 p,q で生成され $pq=1$ という関係のみを満たすモノイド $B=\langle\{p,q\}\mid pq=1\rangle$ (これは *bicyclic* モノイドと呼ばれる) と同型となる．このとき統語準同型 $\eta_{\mathsf{D}}:A^*\to B$ は $\mathrm{Trim}^*(w)=b^ma^n$ となる各 $w\in A^*$ に対して $\eta_{\mathsf{D}}(w)=q^mp^n$ と写し，B の単位元 1 について $\eta_{\mathsf{D}}^{-1}(1)=\mathsf{D}$ が成り立つ．モノイド B においては $(m,n)\neq(m',n')$ ならば $q^mp^n\neq q^{m'}p^{n'}$ のため，特に B は無限モノイドである．

最後に，回文全体の集合 $\mathsf{P}=\{\varepsilon,a,b,aa,bb,aaa,aba,\cdots\}$ について考える．このとき実は $u\equiv_{\mathsf{P}}v\Leftrightarrow u=v$ であることが示せ，統語モノイドは A^* そのものとなる．このように統語的合同関係が自明な恒等関係になる言語は**分離的** (*disjunctive*) と呼ばれ，分離的な言語の統語モノイド (すなわち自由モノイド) には代数的な情報はなにもない．

例 27 で見たように，言語によっては統語モノイドは有限になったり無限になったりするが，以下の定理は正則言語の場合は必ず有限になりまたその逆も成り立つことを言っている代数的言語理論の基本的な定理である．

定理 28 (Myhill-Nerode [16])**.** 任意の言語 $L\subseteq A^*$ について，L が正則であることと，$\mathsf{Synt}(L)$ が有限であることは同値.

ある有限モノイド M に対して，一般に $M\cong\mathsf{Synt}(L)$ となるような正則言語 L は多数存在する (例えば $\mathsf{Synt}((aa)^*)=\mathsf{Synt}(a(aa)^*)\cong\mathbb{Z}/2\mathbb{Z}$，より一般に言語 L とその補集合 $\overline{L}$ の統語モノイドは等しい)．逆に，ある言語 L に対して $L=\eta^{-1}(S)$ を満たすような準同型 $\eta:A^*\to M$ と部分集合 $S\subseteq M$ を持つ有限モノイドは多数ある．このように，有限モノイド全体と正則言語全体の間には多対多の関係があるが,「ある種の良い閉包性を満たす正則言語のクラス全体」と「ある種の良い閉包性を満たす有限モノイドの族全体」には綺麗な 1 対 1 が存在することが知られている．

定義 29 (正則言語の多様体)**.** 正則言語のクラス $\mathcal{C}\subseteq\mathrm{REG}$ が次の 3 つの閉包性を満たすとき，$\mathcal{C}$ を多様体 (*variety*) と呼ぶ.

(L1) $\mathcal{C}$ は Bool 演算に閉じている $(L,M\in\mathcal{C}\Rightarrow\overline{L},L\cup M,L\cap M\in\mathcal{C})$.

(L2) $\mathcal{C}$ は左右からの商に閉じている $(L\in\mathcal{C},a\in A\Rightarrow a^{-1}L,La^{-1}\in\mathcal{C})$.

(L3) $\mathcal{C}$ はモノイド準同型の逆像に閉じている $(L\in\mathcal{C}_A,h:B^*\to A^*\Rightarrow h^{-1}(L)\in\mathcal{C}_B)$.

定義 30 (有限モノイドの多様体)**.** 有限モノイドの族 $\mathsf{V}\subseteq\mathsf{Mon}$ が次の 3 つの閉包性を満たすとき，V を多様体 (*pseudovariety*) と呼ぶ.

(M1) **除モノイド** (*divisor*) を取る操作について閉じている．ここでモノイド M が N の除モノイドであるとは，ある準同型 $h:N\to M$ と N の部分モノイド N' が存在して $h(N')=M$ が成り立つことを言う.

(M2) 有限直積に閉じている $(M_1,\ldots,M_n\in\mathsf{V}\Rightarrow M_1\times\cdots\times M_n\in\mathsf{V})$.

有限モノイドの族 $\mathsf{V}\subseteq\mathsf{Mon}$ を含む最小の多様体を $\langle\mathsf{V}\rangle$ で表す (多様体の積集合も多様体になるため，$\langle\mathsf{V}\rangle$ は常に一意に存在する).

定理 31 (Eilenberg の多様体定理 [10])**.** 正則言語の多様体全体と有限モノイドの多様体全体は一対一に対応する．より精確に，正則言語の多様体 $\mathcal{C} \subseteq \mathrm{REG}$ および有限モノイドの多様体 $\mathsf{V} \subseteq \mathsf{Mon}$ に対する写像

$$F(\mathcal{C}) = \langle \{\mathsf{Synt}(L) \mid A \text{ は文字集合}, L \in \mathcal{C}_A\} \rangle \quad G(\mathsf{V}) = \{L \subseteq A^* \mid \mathsf{Synt}(L) \in \mathsf{V}\}_{A:\text{有限アルファベット}}$$

がそれぞれ正則言語の多様体と有限モノイドの多様体間の全単射となっており，さらに互いに逆写像となっている．

正則言語全体 REG_A は自明な多様体の例であり，これに対応する有限モノイドの多様体は有限モノイド全体 Mon である．歴史的には，非常に多くの非自明な正則言語の多様体と有限モノイドの多様体が発見されており (*cf.* [17, 8])，その中でも古典かつ有名な例は次の星無し言語と非周期的な有限モノイドの対応であろう．ここで，モノイド M が**非周期的** (*aperiodic*) であるとは，任意の $m \in M$ に対してある $n \geq 1$ が存在して $m^n = m^{n+1}$ が成り立つことを言う．

定理 32 (Schützenberger の定理 [19])**.** 星無し言語の多様体 SF と非周期的な有限モノイドの多様体 Ap が対応する．すなわち $L \in \mathrm{SF} \Leftrightarrow \mathsf{Synt}(L) \in \mathsf{Ap}$ が成り立つ．

例 2 で $(aa)^* \subseteq \{a\}^*$ が星無し言語でないことを証明無しで言及したが，上の Schützenberger の定理を使えば即座に示すことができる：$(aa)^*$ の統語モノイドは位数 2 の群 $\mathbb{Z}/2\mathbb{Z}$ であり，単位元ではない元 $x \in \mathbb{Z}/2\mathbb{Z}$ について $x^n \neq x^{n+1}$ が任意の $n \geq 1$ で成り立つため周期的である．

次項からは密度・可測性で定義されるような正則言語の部分族についての考察を展開していくが，一般に，密度や可測性は準同型の逆像について閉じていない．そのため「密度が 0 か 1 となる正則言語全体」のクラス ZO は多様体とはならない．

例 33. L を A 上の正則可測でない言語とする．A に新たな文字 $c \notin A$ を加えたアルファベット $B = A \cup \{c\}$ を考えると，L を B 上の言語と考えると A^* は c を禁句として持つため無限の猿定理より $\delta_B^*(A^*) = 0$ であり，よって $L \subseteq B^*$ の密度は B 上で 0 となり，L は正則言語の定数列 $(\emptyset, A^*)_n$ で明らかに近似できる．そのため L は B においては可測であるが，A^* から B^* への包含準同型 $h : A^* \to B^*$ の逆像 $L = h^{-1}(L)$ は A 上の言語で仮定より可測ではない (よって密度も持たない)．そのため，密度や可測性は準同型の逆像によって保存されない性質であることがわかる．一般に，密度や可測性などの言語の母関数に立脚した概念は準同型の像や逆像と相性が悪い (大抵の場合は保存されない)．

しかし，ZO は準同型の逆像に閉じていないだけで次の意味で多様体を一般化した概念の具体例となっており，さらに一般化された多様体概念でも Eilenberg 型の定理が成り立つことが知られている．

定義 34 (正則言語の局所多様体)**.** 正則言語の族 $\mathcal{C} \subseteq \mathrm{REG}_A$ が定義 29 の 2 つの閉包性 (**L1**) と (**L2**) を満たすとき， $\mathcal{C}$ を A 上の局所多様体 (*local variety*) と呼ぶ．

定義 35 (有限モノイドの局所多様体)**.** アルファベット A で生成される有限モノイドの族 $\mathsf{V} \subseteq \mathsf{Mon}$ が次の閉包性を満たすとき，V を A 上の局所多様体 (*local pseudovariety*) と呼ぶ．

(M'1) 準同型の像を取る操作について閉じている．

(M'2) 部分直積 (*subdirect product*) に閉じている．すなわち $M_1, \ldots, M_n \in \mathsf{V}$ ならば，その直積 $M_1 \times \cdots \times M_n$ の部分モノイド M が「各 $1 \leq i \leq n$ から自然に誘導される射影 $p_i : M \to M_i$ が全て全射」という性質を満たすならば $M \in \mathsf{V}$．

定理 36 (局所多様体版の Eilenberg 型定理 [1])**.** 正則言語の局所多様体全体と有限モノイドの局所多様体全体は一対一に対応する．

局所多様体は、言語クラスとなる多様体とは異なり、固定されたアルファベット A 上の言語族に対する概念である．A 上の正則言語全体 REG_A が局所多様体であることは，REG が多様体であることから明らかであるが，2 節で導入した言語族 FIN と ZO も次で説明するように局所多様体となっており，次項から $\mathrm{FIN}, \mathrm{ZO}, \mathrm{SF}$ のそれぞれの可測性について中心的に議論を行う．

例 37. 有限または補有限な言語全体のクラスを FIN とし，冪零な有限モノイドの全体を Nil とする．ただしモノイド M が**冪零** (*nilpotent*) であるとは，ある元 $z \in M$ が存在して任意の $m \in M$ に対して $z \cdot m = m \cdot z = z$ が成り立ち (この様な元 z を零元と呼ぶ)，任意の m に対してある $i \geq 1$ で $m^i = z$ となることを言う．FIN と Nil はそれぞれ正則言語と有限モノイドの局所多様体となっている．さらに，FIN と Nil が対応する．すなわち $L \in \mathrm{FIN} \Leftrightarrow \mathsf{Synt}(L) \in \mathsf{Nil}$.

例 38 ([20])**.** 密度が 0 か 1 の正則言語全体のクラスを ZO とし，零元を持つ有限モノイド全体を Z とする．ZO と Z はそれぞれ局所多様体となっている．さらに，ZO と Z が対応する．すなわち，$L \in \mathrm{ZO} \Leftrightarrow \mathsf{Synt}(L) \in \mathsf{Z}$.

4.2 局所多様体の閉包作用素としての Carathéodory 拡張

定理 39. アルファベット A 上の言語族 $\mathcal{C} \subseteq 2^{A^*}$ を $\mathcal{C}$ 可測な言語全体へ写す以下の写像

$$\mathrm{Ext}_A(\mathcal{C}) = \{L \subseteq A^* \mid L \text{ は } A \text{ 上 } \mathcal{C} \text{ 可測 }\}$$

は，A 上の言語族の閉包作用素となっている．すなわち，$\mathcal{C}, \mathcal{D} \subseteq 2^{A^*}$ について以下を満たす：

拡大的： $\mathcal{C} \subseteq \mathrm{Ext}_A(\mathcal{C})$.

単調的： $\mathcal{C} \subseteq \mathcal{D}$ ならば $\mathrm{Ext}_A(\mathcal{C}) \subseteq \mathrm{Ext}_A(\mathcal{D})$.

冪等的： $\mathrm{Ext}_A(\mathrm{Ext}_A(\mathcal{C})) = \mathrm{Ext}_A(\mathcal{C})$.

Proof. 拡大的・単調的であることは定義より明らかであるため，冪等的であることを示す．$L \in \mathrm{Ext}_A(\mathrm{Ext}_A(\mathcal{C}))$ について考える．定義より L に両側から収束するような $\mathrm{Ext}_A(\mathcal{C})$ に属す言語の対列 $(K_n, M_n)_n$ が存在する．各 n に対して K_n, M_n はそれぞれ $\mathrm{Ext}_A(\mathcal{C})$ に属するため，K_n に内側から収束する $\mathcal{C}$ の列 $(K_{(n,i)})_i$ と M_n に外側から収束する $\mathcal{C}$ の列 $(M_{(n,i)})_i$ が存在する．このとき，$(K_{(n,n)}, M_{(n,n)})_n$ は L に両側から収束する $\mathcal{C}$ の対列となっているため $L \in \mathrm{Ext}_A(\mathcal{C})$ が成り立つ． □

また，任意の正則言語は密度を持ち (系 9)，さらに正則言語全体は Bool 演算と左右からの商について閉じているため，局所多様体の定義と定理 22–23 から次の系が得られる．

系 40. 正則言語の局所多様体 $\mathcal{C} \subseteq \mathrm{REG}_A$ について，$\mathrm{Ext}_A(\mathcal{C}) \cap \mathrm{REG}_A$ も局所多様体である．

1 節で提起したとおり，拡張 $\mathrm{Ext}_A(\mathrm{REG})$(または $\mathrm{Ext}_A(\mathrm{REG}) \cap \mathrm{CFL}$) の性質を明らかにしたい．しかし，2.3 節で説明したとおり，$\mathrm{Ext}_A(\mathrm{REG})$ は非常に複雑な言語も含んでおり，また CFL は理論的に明らかになっていない点が多く，これは現時点で難しい問題と思われる．

本節では以降，$\mathrm{Ext}_A(\mathrm{REG})$ や $\mathrm{Ext}_A(\mathrm{REG}) \cap \mathrm{CFL}$ のある種の「ミニチュア」である $\mathrm{Ext}_A(\mathrm{ZO}) \cap \mathrm{REG}$ や $\mathrm{Ext}_A(\mathrm{SF}) \cap \mathrm{REG}$ について考察する．以降，拡張の像を正則言語に制限した写像を $\mathrm{REext}_A(\mathcal{C}) = \mathrm{Ext}_A(\mathcal{C}) \cap \mathrm{REG}_A$ で表す．

定義より，密度が 0 (あるいは 1) な言語の列 $(L_n)_n$ で近似できるような言語は同じく密度が 0 (あるいは 1) な言語に限る．よって次の定理が成り立つ ($\mathrm{Ext}_A(\mathrm{ZO}_A)$ は非正則な言語も含むため $\mathrm{ZO}_A \subsetneq \mathrm{Ext}_A(\mathrm{ZO}_A)$ となることに注意)．また，似たような議論で有限・補有限な局所多様体 FIN_A についても，REext_A は拡張しないことが示せる．

定理 41. 任意の A について，$\mathrm{REext}_A(\mathrm{ZO}_A) = \mathrm{ZO}_A$.

定理 42. 任意の A について，$\mathrm{REext}_A(\mathrm{FIN}_A) = \mathrm{FIN}_A$.

また，アルファベットが単元集合からなる場合は，$\mathrm{SF}_A = \mathrm{FIN}_A$ が成り立つことが知られているため，次の系が得られる．

系 43. アルファベット $A = \{a\}$ においては $\mathrm{Ext}_A(\mathrm{SF}_A) = \mathrm{SF}_A = \mathrm{FIN}_A$ が成り立つ．

しかし，アルファベットが単元集合でない場合は事情が異なってくる．

定理 44. 2 つ以上の文字を含む A において，$\mathrm{REext}_A(\mathrm{SF}_A) \supsetneq \mathrm{SF}_A$.

Proof. 前項で説明したとおり $(aa)^* \notin \mathrm{SF}_A$ であるため，$(aa)^* \in \mathrm{REext}_A(\mathrm{SF}_A)$ であることを示す．A は 2 つ以上の文字を含むため，$(aa)^* \subseteq a^* = A^* \setminus A^*(A \setminus \{a\})A^*$ の密度は共に 0 となり，したがって星無し言語の対の定数列 $(\emptyset, a^*)_n$ が $(aa)^*$ に両側から収束する． □

定理 45. アルファベット A 上の星無し言語 $L \in \mathrm{SF}_A$ について，$\delta^*_A(L) > 0$ ならば L は奇数長の語も偶数長の語も両方含む．

Proof. $\eta_L : A^* \to \mathsf{Synt}(L)$ を L の統語準同型とし，$\eta_L(L) = S \subseteq \mathsf{Synt}(L)$ と置く．L は正則言語であるため $\mathsf{Synt}(L)$ は有限である．モノイド M の空でない部分集合 $I \neq \emptyset$ が $MIM = \{abc \mid a, c \in M, b \in I\} \subseteq I$ を満たすとき I をイデアルという．イデアル I と J の積 $IJ \subseteq I \cap J$ もまたイデアルになるため，イデアル全体の積は M が有限な場合は (包含関係で) 最小なイデアルとなる (M が無限の場合は全体の積が空集合になり得る)．$\mathsf{Synt}(L)$ の最小イデアルを K と置き，$k \in K$ の逆像となる語をそれぞれ $w_k \in \eta_L^{-1}(k)$ と置く (η_L は全射なので必ずそのような語が存在する)．

このとき，$\delta^*_A(L) > 0$ という仮定から $S \cap K \neq \emptyset$ が言える．なぜなら，K に属さない任意の元 $m \in \mathsf{Synt}(L) \setminus K$ および任意の文字列 $x, y \in A^*$ について，$\eta_L(xw_ky) = \eta_L(x) \cdot k \cdot \eta_L(y) \in K$ のため特に $\eta_L(xw_ky) \neq m$ となるからである．これはすなわち $\eta_L^{-1}(m)$ が w_k を禁句に持つことを意味しており，よって無限の猿定理より $\eta_L^{-1}(m)$ の密度は 0 となる．密度が 0 な言語の有限個の和集合はやはり密度 0 であるため，$\delta^*_A(\eta_L^{-1}(K)) = 1$ が成り立つ．そのため，L の密度が非零であるためには S が少なくとも 1 つ K の元を含んでいる必要がある．その様な元 $t \in S \cap K$ を 1 つ固定する．

$\delta^*_A(\eta_L^{-1}(K)) = 1$ であるため，特に $\eta_L^{-1}(K)$ はある奇数長の語 w_{odd} を必ず含む (そうでないとすると密度が 1/2 以下になってしまう)．w_{odd} の像を $m_{\mathrm{odd}} = \eta_L(w_{\mathrm{odd}})$ とする．このとき，Schützenberger の定理より $\mathsf{Synt}(L)$ は非周期的であるため，ある $i \geq 1$ が存在して $m^i_{\mathrm{odd}} = m^{i+1}_{\mathrm{odd}}$ が成り立つ．K は最小イデアルであるため，$x \cdot m^i_{\mathrm{odd}} \cdot y = t$ となる元 $x, y \in \mathsf{Synt}(L)$ が存在する (そうでないとすると m_{odd} で生成されるイデアル $\mathsf{Synt}(L) \cdot m_{\mathrm{odd}} \cdot \mathsf{Synt}(L)$ が K よりも真に小さいイデアルとなってしまうため K の最小性に反する)．そのような元 x, y に対して $w_x \in \eta_L^{-1}(x), w_y \in \eta_L^{-1}(y)$ をそれぞれ固定する．すると 2 つの語 $w_x w^i_{\mathrm{odd}} w_y$ と $w_x w^{i+1}_{\mathrm{odd}} w_y$ は

$$\eta_L(w_x w^i_{\mathrm{odd}} w_y) = x \cdot m^i_{\mathrm{odd}} \cdot y = t = x \cdot m^{i+1}_{\mathrm{odd}} \cdot y = \eta_L(w_x w^{i+1}_{\mathrm{odd}} w_y)$$

を満たすため，それぞれ L に属する．w_{odd} の長さは奇数であったため，$w_x w^i_{\mathrm{odd}} w_y$ と $w_x w^{i+1}_{\mathrm{odd}} w_y$ の長さは偶奇が異なる．よって定理が示された． □

言語 $(AA)^*$ の密度は 1/2 であるが，定理 45 より正測度な星無し言語 L は必ず奇数長の語を含むため $L \subsetneq (AA)^*$ となり，したがって $\underline{\mu}_{\mathrm{SF}_A}(AA^*) = 0$ となり，補集合について考えると同様に $\overline{\mu}_{\mathrm{SF}_A}(AA^*) = 1$ が得られる．

系 46. 任意のアルファベット A において $(AA)^* \notin \mathrm{Ext}_A(\mathrm{SF}_A)$. 特に，$(AA)^*$ の SF ギャップは 1.

定理 44 と系 46 をまとめると，REext_A は SF を非自明な局所多様体に拡張することがわかる：

系 47. 2 つ以上の文字を含む A において，$\mathrm{SF}_A \subsetneq \mathrm{REext}_A(\mathrm{SF}_A) \subsetneq \mathrm{REG}_A$.

定理 48. A 上の正則言語 L について，以下は同値：

1. L は SF_A 可測.
2. $\mathsf{Synt}(L)$ の最小イデアルが非自明な部分群を含まない.

5　おわりに

本論文では可測性の一般的な性質や，いくつかの局所多様体に対して可測性を明らかにしてきた．1 節で提起したとおり，拡張 $\mathrm{Ext}_A(\mathrm{REG})$(または $\mathrm{RExt}_A(\mathrm{REG}) \cap \mathrm{CFL}$) の性質を明らかにしたいといのが本研究の基本的な動機である．しかし，現時点では文脈自由言語は明らかになっていない点が多く，$\mathrm{Ext}_A(\mathrm{REG}) \cap \mathrm{CFL}$ の考察は理論的に興味深いものの難しい問題と思われる．例えば予想 24 のような基本的な問題も未解決であり，また次の問題も未解決である．無曖昧文脈自由言語に対しては代数関数の理論を使うことで密度を持つことが言えるはず (あるいは folklore) であるが，一般の文脈自由言語については母関数の理論が確立していないため別の道具が必要かもしれない．

問題 49. 任意の文脈自由言語は密度を持つか？

前節では $\#(A) \geq 2$ の場合，RExt_A は SF_A を非自明な局所多様体に拡張したが，この局所多様体の特徴づけは今後の課題である．

問題 50. $\mathrm{RExt}_A(\mathrm{SF}_A)$ はどのような局所多様体か？$\mathrm{RExt}_A(\mathrm{SF}_A)$ に対応する有限モノイドの局所多様体は何か？正則言語 L について $L \in \mathrm{RExt}_A(\mathrm{SF}_A)$ を決定するアルゴリズムは存在するか？

また，Eilenberg 型の定理 (定理 36) から，A 上の正則言語の局所多様体 $\mathcal{C}$ には A 生成有限モノイドの局所多様体 $\mathsf{V}(\mathcal{C})$ が，逆に A 生成有限モノイドの局所多様体 V には A 上の正則言語の局所多様体 $\mathcal{C}(\mathsf{V})$ が一対一に対応する．よって正則言語の局所多様体上の閉包作用素 RExt_A に対応する有限モノイドの局所多様体上の閉包作用素 $\mathrm{MExt}_A : \mathsf{V} \mapsto \mathsf{V}(\mathrm{RExt}_A(\mathcal{C}(\mathsf{V})))$ が自然に誘導される．

問題 51. 有限モノイドの局所多様体の閉包作用素 MExt_A を純代数的に特徴づけられるか？

言語 L が正則可測であるとは直感的には「L が正則言語で任意の精度で近似できる」ということであり，自然な概念である．計算量を下げるために決定問題の厳密解ではなく近似解を求めることはよくあり，正則可測性がそのような場面に理論的・実用的に応用できる可能性もあるかもしれない．また，言語の学習においてはある種の収束概念 (極限同定) を考えるが，これを正則可測性の文脈での収束に置き換えるとどうなるだろうか．正則可測性の応用は今後の課題であり，本論文を通じて正則可測性を知った読者の方々にもぜひ考えてもらえるとありがたい．

謝辞　丁寧に論文を読んでコメントいただいた PPL2021 の査読者に感謝する．本研究は JSPS 科研費 JP19K14582 の助成を受けたものである．

参考文献

[1] Adámek, J., Milius, S., Myers, R. S. R., and Urbat, H.: Generalized Eilenberg Theorem I: Local Varieties of Languages, *Foundations of Software Science and Computation Structures*, Muscholl, A.(ed.), Berlin, Heidelberg, Springer Berlin Heidelberg, 2014, pp. 366–380.

[2] Berstel, J.: Sur la densité asymptotique de langages formels, *International Colloquium on Automata, Languages and Programming*, France, North-Holland, 1973, pp. 345–358.

[3] Berstel, J. and Perrin, D.: *Theory of codes*, Pure and Applied Mathematics, Vol. 117, Academic Press Inc., 1985.

[4] Berstel, J., Perrin, D., and Reutenauer, C.: *Codes and Automata*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2009.

[5] Buck, R. C.: The Measure Theoretic Approach to Density, *American Journal of Mathematics*, Vol. 68, No. 4(1946), pp. 560–580.

[6] Chomsky, N. and Schützenberger, M.: The Algebraic Theory of Context-Free Languages*, *Computer Programming and Formal Systems*, Vol. 35, Elsevier, 1963, pp. 118–161.

[7] Dömösi, P. and Ito, M.: *Context-Free Languages and Primitive Words*, World Scientific Publishing Company, 2014.

[8] Diekert, V., Gastin, P., and Kufleitner, M.: A Survey on Small Fragments of First-Order Logic over Finite Words, *Int. J. Found. Comput. Sci.*, Vol. 19, No. 3(2008), pp. 513–548.

[9] Dömösi, P., Horváth, S., and Ito, M.: On the connection between formal languages and primitive words, 1991, pp. 59–67.

[10] Eilenberg, S. and Tilson, B.: *Automata, languages and machines. Volume B*, Pure and applied mathematics, Academic Press, New-York, San Franciso, London, 1976.

[11] Eisman, G. and Ravikumar, B.: Approximate Recognition of Non-regular Languages by Finite Automata, *Twenty-Eighth Australasian Computer Science Conference (ACSC2005)*, CRPIT, Vol. 38, Newcastle, Australia, ACS, 2005, pp. 219–228.

[12] Flajolet, P.: Ambiguity and transcendence, *Automata, Languages and Programming*, Berlin, Heidelberg, Springer Berlin Heidelberg, 1985, pp. 179–188.

[13] Flajolet, P.: Analytic models and ambiguity of context-free languages, *Theoretical Computer Science*, Vol. 49, No. 2(1987), pp. 283–309.

[14] Greibach, S. A.: A note on undecidable properties of formal languages, *Mathematical systems theory*, Vol. 2(1968), pp. 1–6.

[15] Kemp, R.: A note on the density of inherently ambiguous context-free languages, *Acta Informatica*, Vol. 14, No. 3(1980), pp. 295–298.

[16] Myhill, J. R.: Finite Automata and the Representation of Events, Technical Report WADC TR-57-624, Wright-Paterson Air Force Base, 1957.

[17] Pin, J.-E.: Mathematical Foundations of Automata Theory, 2012.

[18] Salomaa, A. and Soittola, M.: *Automata Theoretic Aspects of Formal Power Series*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1978.

[19] Schützenberger, M.-P.: On finite monoids having only trivial subgroups, *Information and Control*, Vol. 8, No. 2(1965), pp. 190–194.

[20] 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*, 2015, pp. 172–185.

[21] Sin'ya, R.: Asymptotic Approximation by Regular Languages, Proceedings of the 47th International Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'21), to appear, 2021.

[22] 新井仁之: ルベーグ積分講義—ルベーグ積分と面積 *0* の不思議な図形たち, 日本評論社, 2003.

[23] 新屋良磨: オートマトン理論再考, コンピュータ ソフトウェア, Vol. 34, No. 3(2017), pp. 3–35.