Carathéodory Extensions of Subclasses of Regular Languages

Ryoma Sin'ya

Akita University
ryoma@math.akita-u.ac.jp

Abstract. A language L is said to be regular measurable if there exists an infinite sequence of regular languages that "converges" to L. In [1], the author showed that, while many complex context-free languages are regular measurable, the set of all primitive words and certain deterministic context-free languages are regular *im*measurable. This paper investigates general properties of measurability, including closure properties, decidability and different characterisation. Further, for a suitable subclass $\mathcal{C}$ of regular languages, we show that the class of all $\mathcal{C}$-measurable regular languages has a good algebraic structure.

1 Introduction

How can we measure the volume of an object with a very complex shape? If it can be wet, an easy way is to slowly and completely submerge the object suspended by a thread in a rectangular tank filed with water, pull it out, and calculate the amount of water that overflows from the reduced water level. The amount of water that overflows is needed to "cover" the object, so it will be a good estimation of the volume of the object. It is a standard way in measure theory to cover an object $X \subseteq \mathbb{R}^d$ with a set $Y \supseteq X$ with good properties, called a "basic set", and use the measure of Y as an estimation (from outer) of the measure of X.

For example, in the case of Lebesgue measure (*cf.* [2]), we define the length of an interval $I = [a,b], [a,b), (a,b], (a,b)$ as $|I| = b - a$, and call the direct product $B = I_1 \times \cdots \times I_d$ of d intervals as a box (with $|B| = |I_1| \times \cdots \times |I_d|$), and *regard a countable union of boxes as a basic set.* The Lebesgue outer measure of a set $X \subseteq \mathbb{R}$ is defined as

$$m^*(X) = \inf \left\{ \sum_{n=1}^{\infty} |B_n| \;\middle|\; \bigcup_{n=1}^{\infty} B_n \supseteq X; B_n \text{ is a box for each } n \geq 1 \right\},$$

i.e., the lower bound on the volume required to cover X by a basic set $\bigcup_{n=1}^{\infty} B_n$. X is said to be Lebesgue measurable if it satisfies the following so-called *Carathéodory's condition* (where $\overline{X}$ is the complement of X):

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

Actually, for subsets of the set of natural numbers $\mathbb{N}\,(\ni 0)$, we can apply this measure theoretic approach. In [3], Buck defines the density of an arithmetic progression (AP for short) $A = \{pn + q \mid n \in \mathbb{N}\}$ where $p, q \in \mathbb{N}$ [1] as $d(A) = 1/p$ ($d(A) = 0$ if $p = 0$), *regards a finite union of arithmetic progressions as a basic set,* and defines the outer density of $X \subseteq \mathbb{N}$ as

$$d^*(X) = \inf\left\{ \sum_{n=1}^{k} d(A_n) \;\middle|\; \bigcup_{n=1}^{k} A_n \supseteq X; k \in \mathbb{N}, A_n \text{ is an AP for each } n \in [1,k] \right\}.$$

As like the Lebesgue measurability, $X \subseteq \mathbb{N}$ is said to be measurable if it satisfies the Carathéodory's condition

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

and Buck called $d^*(X)$ *the measure density of* X in this case.

Regular measurability (REG-measurability) proposed in [1] is an adoption of the Buck's measure density for formal languages: we define the density of a language $L \subseteq A^*$, *regards a regular language as a basic set*, and define the measurability of a language via outer and inner density (precise definition appears in the next section). The main motivation of [1] is, not just to generalise Buck's measure density, but also to tackle a long-standing open problem so-called *primitive words conjecture*. Some non-trivial partial results can be found in [1], which we will briefly describe in the next section. Regular measurability is an emergent notion and hence its theory is not well developed yet. In fact, it is fair to say that very little is known about the class of all regular measurable languages (regular measurable context-free languages, respectively). This paper investigates fundamental properties of regular measurability (and $\mathcal{C}$-measurability for a general language class $\mathcal{C}$) like as closure properties, decidability and different characterisation. Moreover, as a "miniature" of regular measurability, for some subclass $\mathcal{C}$ of regular languages, we investigate $\mathcal{C}$-measurability. While the class of all regular measurable languages (regular measurable context-free languages) has a complex structure and it is somewhat hard to analyse, for some suitable subclass $\mathcal{C}$ (called *local variety*) of regular languages, we will show that the class of all $\mathcal{C}$-measurable regular languages has a good algebraic structure and more easier to analyse.

Our contribution and the organisation of the paper

In this paper, all theorems/corollaries without citation are new (as much as we know), and main results consist of three kinds: **(I)** Give some new examples of regular (im)measurable languages (Theorem 5–6, Corollary 1 in Section 2). **(II)** Show some closure properties, an undecidability result (modulo a certain conjecture), and a different characterisation via the Carathéodory's condition of $\mathcal{C}$-measurability for a general language class $\mathcal{C}$ (Theorem 7–10 in Section 3). **(III)** Examine Carathéodory extensions of some local varieties of regular languages (Theorem 13–16 in Section 4). We also discuss future work and pose few open problems in Section 5.

[1] Here p can be 0 and we call a singleton $\{q\}$ arithmetic progression in this case.

2 Density and Measurability

This section provides the precise definitions of density and measurability. In Section 2.3, we briefly describe results in [1], and also give some new examples of regular measurable/immeasurable languages.

2.1 Density of formal languages

For a set X, we denote by $\#(X)$ the cardinality of X. We denote by $\mathbb{N}$ the set of natural numbers including 0. For an alphabet A, we denote the set of all words (all non-empty words, respectively) over A by A^* (A^+, respectively). For a word $w \in A^*$ and a letter $a \in A$, $|w|_a$ denotes the number of occurrences of a in w. A word v is said to be a subword of a word w if $w = xvy$ for some $x, y \in A^*$. For a language $L \subseteq A^*$, we denote by $\overline{L} = A^* \setminus L$ the complement of L. We say that L is co-finite if its complement is finite. A language L is said to be *dense* if $L \cap A^*wA^* \neq \emptyset$ holds for any $w \in A^*$. L is not dense means $L \cap A^*wA^* = \emptyset$ for some word w by definition, and such word is called a forbidden word of L.

Definition 1. Let $L \subseteq A^*$ be a language. The *density* $\delta_A^*(L)$ of L over A is defined as

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

if its exists, otherwise we write $\delta_A^*(L) = \perp$ and say that L *does not have a density*. L is called *null* if $\delta_A^*(L) = 0$, and conversely L is called *co-null* if $\delta_A^*(L) = 1$.

The following observation is basic. See Chapter 13 of [4] for more details.

Lemma 1. *Let $K, L \subseteq A^*$ with $\delta_A^*(K) = \alpha, \delta_A^*(L) = \beta$. Then we have:*

(1) $\alpha \leq \beta$ *if* $K \subseteq L$.
(2) $\delta_A^*(L \setminus K) = \beta - \alpha$ *if* $K \subseteq L$. *In particular,* $\delta_A^*(\overline{K}) = \delta_A^*(A^* \setminus K) = 1 - \alpha$.
(3) $\delta_A^*(K \cup L) \leq \alpha + \beta$ *if* $\delta_A^*(K \cup L) \neq \perp$.
(4) $\delta_A^*(K \cup L) = \alpha + \beta$ *if* $K \cap L = \emptyset$.
(5) $\delta_A^*(wK) = \delta_A^*(Kw) = \alpha/\#(A)^{|w|}$ *for each* $w \in A^+$.

Example 1. Here we enumerate a few examples of densities of languages.

(1) Consider $(AA)^*$ the set of all words with even length. Because $\frac{\#((AA)^* \cap A^n)}{\#(A^n)}$ is 1 if n is even otherwise 0, clearly $\delta_A^*((AA)^*) = 1/2$ holds.
(2) For each word w, the language A^*wA^*, *i.e.*, the set of all words that contain w as a subword, has density 1 (co-null). This fact is sometimes called *infinite monkey theorem*. A language L having a forbidden word w is always null; since $A^*wA^* \subseteq \overline{L}$ holds by definition, we have $\delta_A^*(A^*wA^*) \leq \delta_A^*(\overline{L})$ which implies $\delta_A^*(\overline{L}) = 1$ by infinite monkey theorem. Thus L is null.

(3) The following language

$$L_\perp = \{w \in A^* \mid 3^n \leq |w| < 3^{n+1} \text{ for some even number } n\}$$

does not have a density ($\delta_A^*(L_\perp) = \perp$). The proof is as follows. The density of $L_\perp$ is the limit (when $n \to \infty$) of the fraction

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

if it exists by definition. Consider $n = 3^k$ for some even number k and let $0 \leq \alpha \leq 1$ be the value of the fraction (1) of $n = 3^k$, then the value (1) of $n = 3^{k+1}$ satisfies

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

Conversely, consider $n = 3^k$ for some odd number k and let $0 \leq \beta \leq 1$ be the value (1) of $n = 3^k$, then the value (1) of $n = 3^{k+1}$ satisfies

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

Hence the value (1) could be larger than 2/3 and smaller than 1/3 infinitely many times so that $\delta_A^*(L_\perp)$ diverges. □

Example 1 shows us that, for some language, its density is either zero or one, for some, like $(AA)^*$, a density could be a rational number like 1/2, and for some, like $L_\perp$ a density may not even exist. However, the following theorem tells us that all regular languages *do* have densities.

Theorem 1 (*cf.* Theorem III.6.1 of [5]). *Every regular language has a density and it is rational.*

Also, for the class of regular languages, two notions "not null" (a measure theoretic largeness) and "dense" (a topological largeness) are equivalent.

Theorem 2 ([6]). *A regular language L is not null if and only if L is dense.*

2.2 $\mathcal{C}$-measurability of formal languages

A *language class* $\mathcal{C}$ is a family of languages $\{\mathcal{C}_A\}_{A:\text{ finite alphabet}}$ where $\mathcal{C}_A \subseteq 2^{A^*}$ for each A and $\mathcal{C}_A \subseteq \mathcal{C}_B$ for each $A \subseteq B$. We simply write $L \in \mathcal{C}$ if $L \in \mathcal{C}_A$ for some alphabet A. We denote by REG and CFL the class of regular languages and context-free languages, respectively.

We now introduce the notion of $\mathcal{C}$-measurability which is a formal language theoretic analogue of Buck's measure density [3].

Definition 2 ([1]). Let $\mathcal{C}$ be a class of languages. For a language $L \subseteq A^*$, we define its *$\mathcal{C}$-inner-density* $\underline{\mu}_{\mathcal{C}_A}(L)$ and *$\mathcal{C}$-outer-density* $\overline{\mu}_{\mathcal{C}_A}(L)$ over A as

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

A language L is said to be *$\mathcal{C}$-measurable* over A if $\underline{\mu}_{\mathcal{C}_A}(L) = \overline{\mu}_{\mathcal{C}_A}(L)$ holds, and we simply write $\overline{\mu}_{\mathcal{C}_A}(L)$ as $\mu_{\mathcal{C}_A}(L)$ in this case. We say that an infinite sequence $(L_n)_n$ of languages over A *converges to L from inner (from outer, respectively)* if $L_n \subseteq L$ ($L_n \supseteq L$, respectively) for each n and $\lim_{n\to\infty} \delta_A^*(L_n) = \delta_A^*(L)$.

Remark 1. Both density and $\mathcal{C}$-measurability depends on the alphabet. For example, any language $L \subseteq A^*$ is of density *zero* over $B \supsetneq A$. Also, any language $L \subseteq A^*$ is REG-measurable over $B \supsetneq A$: clearly $\emptyset \subseteq L \subseteq (B \setminus \{b\})^*$ holds for $b \in (B \setminus A)$ and hence $\underline{\mu}_{\mathrm{REG}_B}(L) = \overline{\mu}_{\mathrm{REG}_B}(L) = 0$ ($(B \setminus \{b\})^*$ has a forbidden word b hence it is null over B by infinite monkey theorem), *i.e.*REG-measurable over B. Hereafter, we mainly consider density and $\mathcal{C}$-measurability over the *minimum* alphabet for each language L, *i.e.*, the minimum alphabet A satisfying $L \subseteq A^*$. We sometimes omit the subscript of $\underline{\mu}_{\mathrm{REG}_A}(L), \overline{\mu}_{\mathrm{REG}_A}(L)$ like $\underline{\mu}_{\mathrm{REG}}(L), \overline{\mu}_{\mathrm{REG}}(L)$, and we simply say "$L$ is of density one" or "L is $\mathcal{C}$-measurable". In this case the considered alphabet is always the minimum one.

The following basic lemmata will be used in the next section.

Lemma 2 (*cf.* [1]). *Let $K, L \subseteq A^*$ be two languages.*

(1) $\underline{\mu}_{\mathcal{C}_A}(K) \leq \delta_A^*(K) \leq \overline{\mu}_{\mathcal{C}_A}(K)$ *if* $\delta_A^*(K) \neq \bot$. *In particular,* $\delta_A^*(K) = \bot$ *implies K is $\mathcal{C}$-immeasurable.*
(2) $\overline{\mu}_{\mathcal{C}_A}(K) \leq \overline{\mu}_{\mathcal{C}_A}(L)$ *if* $K \subseteq L$.
(3) $\overline{\mu}_{\mathcal{C}_A}(K \cup L) \leq \overline{\mu}_{\mathcal{C}_A}(K) + \overline{\mu}_{\mathcal{C}_A}(L)$ *if $\mathcal{C}$ is closed under union.*
(4) $\overline{\mu}_{\mathcal{C}_A}(K) = \delta_A^*(K)$ *if* $K \in \mathcal{C}$ *and* $\delta_A^*(K) \neq \bot$.

Lemma 3 (*cf.* [1]). *Let $\mathcal{C}$ be a language class closed under complementation. A language $L \subseteq A^*$ is $\mathcal{C}$-measurable if and only if*

$$\overline{\mu}_{\mathcal{C}_A}(L) + \overline{\mu}_{\mathcal{C}_A}(\overline{L}) = 1. \tag{2}$$

2.3 Examples of REG-measurable/immeasurable languages

In this subsection we describe several examples of REG-(im)measurable languages. In [1], it is shown that many complex context-free languages are still REG-measurable summarised as follows.

Theorem 3 ([1]). *Following context-free languages are all non-regular but* REG-*measurable.*

(1) $\mathsf{D} = \{w \in \{a,b\}^* \mid |w|_a = |w|_b, |u|_a \geq |u|_b$ *for every prefix u of w*$\}$.
(2) $\mathsf{P} = \{w \in \{a,b\}^* \mid w$ *is a palindrome*$\}$.

(3) $\mathsf{O}_3 = \{w \in \{a,b,c\}^* \mid |w|_a = |w|_b \text{ or } |w|_a = |w|_c\}$.
(4) $\mathsf{O}_4 = \{w \in \{x,\overline{x},y,\overline{y}\}^* \mid |w|_x = |w|_{\overline{x}} \text{ or } |w|_y = |w|_{\overline{y}}\}$.
(5) $\mathsf{G} = \{a^{n_1}ba^{n_2}b\cdots a^{n_k}b \mid k \geq 1, n_i \neq i \text{ for each } 1 \leq i \leq k\}$.
(6) $\mathsf{K} = S_1\{c\}A^* \cup S_2\{c\}A^*$ *where* $A = \{a,b,c\}$ *and*

$$S_1 = \{a\}\{b^ia^i \mid i \geq 1\}^* \qquad S_2 = \{a^ib^{2i} \mid i \geq 1\}^*\{a\}^+.$$

The semi-Dyck language D and the palindromes P are classical examples of non-regular languages. Flajolet [7] showed that the language O_3 and O_4 are inherently ambiguous. Flajolet also showed that the generating function of G is not algebraic [8] and thus it is an inherently ambiguous context-free language due to the well-known Chomsky–Schützenberger theorem [9] stating that the generating function of every unambiguous context-free language is algebraic. The language K defined by Kemp [10] is the first example of a context-free language with *transcendental density*. While several complex context-free languages like G or K are REG-measurable, the following theorem says certain deterministic context-free languages and the set of all primitive words are REG-immeasurable.

Theorem 4 ([1]). *The following languages over* $A = \{a,b\}$ *are all* REG*-immeasurable.*

(1) $\mathsf{M}_n = \{w \in A^* \mid |w|_a > n \cdot |w|_b\}$ *for each* $n \geq 1$.
(2) *The set* Q *of all primitive words. Here* $w \in A^+$ *is said to be primitive if it can not be represented as a power of any shorter words, i.e., for every* $v \in A^+$ *and* $n \in \mathbb{N}$, $v^n = w$ *implies* $v = w$ *and* $n = 1$.

In particular, the above languages are REG*-immeasurable in a strong sense as follows:* $\underline{\mu}_{\mathrm{REG}}(\mathsf{M}_n) = \underline{\mu}_{\mathrm{REG}}(\mathsf{Q}) = 0$ *and* $\overline{\mu}_{\mathrm{REG}}(\mathsf{M}_n) = \overline{\mu}_{\mathrm{REG}}(\mathsf{Q}) = 1$ *for each* $n \geq 2$.

In [1] the author originally conjectured that there is no context-free language like Q: if a context-free language L is co-null, then it can be somehow "approximated" by regular languages from inner, *i.e.*, $\underline{\mu}_{\mathrm{REG}}(L) > 0$. If this conjecture *was* true, then the primitive words conjecture "$\overline{\mathsf{Q}}$ is not context-free" posed by Dömösi, Horváth and Ito [11] was true, too. However, the author found a counterexample $\overline{\mathsf{M}}_2$ and hence this naïve approach did not work (still, this approach has some possibility, see the last section of [1] for details).

Now we give three new examples of REG-(im)measurable languages. The following indexed language is not context-free, but REG-measurable.

Theorem 5. $L_{\exp} = \{a^{2^n} \mid n \in \mathbb{N}\}$ *is* REG*-measurable over* $A = \{a\}$.

Proof. Clearly, $\delta^*_A(L_{\exp}) = 0$ holds hence it is enough to construct a sequence of regular languages that converges to $L_{\exp}$ from outer. For each $k \geq 1$, a regular language $L_k = (a^k)^* \cup \{a^n \mid 0 < n < k\}$ satisfies $\delta^*_A(L_k) = 1/k$ ($\lim_{k\to\infty} \delta^*_A(L_k) = 0$, in particular). We show that $a^{2^n} \in L_{2^k}$ holds for each $k \geq 1$ and $n \in \mathbb{N}$ (*i.e.*, $L_{\exp} \subseteq L_{2^k}$). The case $2^n < 2^k$ is clear by definition thus consider the case $2^n \geq 2^k$. In this case, $2^n = 2^k \cdot 2^{n-k}$ holds hence a^{2^n} is the 2^{n-k} times repetition of a^{2^k} which means $a^{2^n} \in (a^{2^k})^* \subseteq L_{2^k}$. Thus the sequence $(L_{2^k})_{k\geq 1}$ converges to $L_{\exp}$ from outer. □

The next theorem tells us that REG-measurable languages exist for each real number between 0 and 1.

Theorem 6. *Let A be an alphabet including at least two letters. For each real number $0 \leq \alpha \leq 1$, there exists a* REG*-measurable language L over A with density exactly α.*

Proof. Proof. Consider the case $A = \{a, b\}$ (a general case can be shown similarly). Let $(\alpha_n)_{n\geq 1}$ (where each $\alpha_i \in \{0, 1\}$) be the binary expansion of $\alpha \in [0, 1]$: $\alpha = \sum_{n=1}^{\infty} \alpha_n 2^{-n}$. Define $K_0 = \emptyset, M_0 = A^*$ and define K_n, M_n inductively as follows:

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

Clearly, K_n and M_n are regular and $K_n \subseteq K_{n+1} \subseteq M_{n+1} \subseteq M_n$ holds for each n. We have $b^{n-1}aA^* \cap b^{m-1}aA^* = \emptyset$ for each $n \neq m$, and $\delta_A^*(b^{n-1}aA^*) = 2^{-n}$ by Lemma 1-(5). Hence by using Lemma 1-(2) and (2) we can conclude that

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

holds. Thus $(K_n, M_n)_n$ converges to the limit language $L = \bigcup_{n\in\mathbb{N}} K_n = \bigcap_{n\in\mathbb{N}} M_n$ whose density is $\delta_A^*(L) = \lim_{n\to\infty} \delta_A^*(K_n) = \lim_{n\to\infty} \delta_A^*(M_n) = \alpha$. □

Finally, by Lemma 2-(1) we have the following REG-immeasurable language.

Corollary 1. *$L_\perp$ defined in Example 1-*(3) *is* REG*-immeasurable.*

3 Closure Properties and Carathéodory's Condition

In this section we investigate general properties of $\mathcal{C}$-measurability. First we show that $\mathcal{C}$-measurability is closed under Boolean operations and left-and-right quotients, with some density condition. This fact plays important role in the next section.

Theorem 7. *Let $\mathcal{C}$ be a language class closed under Boolean operations. If L, K are $\mathcal{C}$-measurable, and if every language obtained by a finite Boolean combination of languages in $\mathcal{C} \cup \{L, K\}$ has a density, then the complement $\overline{L}$, the union $L \cup K$ and the intersection $L \cap K$ are also $\mathcal{C}$-measurable.*

Proof. $\mathcal{C}$ is closed under Boolean operations by assumption, thus a language L is $\mathcal{C}$-measurable if and only if $\overline{L}$ is $\mathcal{C}$-measurable by Lemma 3. Thus it is enough to show that $L \cup K$ is $\mathcal{C}$-measurable if L and K are. Let $(L_n)_n$ and $(K_n)_n$ are convergent sequence from inner to L and K, respectively. Since $\mathcal{C}$ is closed under taking union, $(L_n \cup K_n)_n$ is a sequence of languages in $\mathcal{C}$. We show that this sequence converges to $L \cup K$ from inner. For any $\epsilon/2 > 0$ there exists δ such that

$\delta^*_A(L) - \delta^*_A(L_n), \delta^*_A(K) - \delta^*_A(K_n) < \epsilon/2$ for each $n > \delta$, thus by Lemma 1-(2) we have

$$(\delta^*_A(L) - \delta^*_A(L_n)) + (\delta^*_A(K) - \delta^*_A(K_n)) = \delta^*_A(L \setminus L_n) + \delta^*_A(K \setminus K_n) < \epsilon.$$

By assumption, $(L \setminus L_n) \cup (K \setminus K_n)$ has a density, hence by the subadditivity of δ^*_A (Lemma 1-(3))

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

holds. Clearly, $(L \cup K) \setminus (L_n \cup K_n) \subseteq (L \setminus L_n) \cup (K \setminus K_n)$ holds and hence $\delta^*_A((L \cup K) \setminus (L_n \cup K_n)) < \epsilon$, which means that $(L_n \cup K_n)_n$ converges to $L \cup K$ from inner. We can also construct an convergent sequence to $L \cup K$ from outer in the same way. □

Theorem 8. *Let $\mathcal{C}$ be a language class closed under left quotients (right quotients, respectively). If L is $\mathcal{C}$-measurable, and if the left quotient $a^{-1}L$ (the right quotient La^{-1}, respectively) has a density, then it is also $\mathcal{C}$-measurable.*

Proof. For a $\mathcal{C}$-measurable language L over A, we show that $a^{-1}L$ is also $\mathcal{C}$-measurable (La^{-1} can be shown by the same way). By definition, there is a convergent sequence $(K_n, M_n)_n$ to L. We show that $(a^{-1}K_n, a^{-1}M_n)_n$ converges to $a^{-1}L$.

For simplicity, we consider the case $A = \{a, b\}$ (a general case can be shown similarly). For each $a \in A$ we have $L \cap aA^* = aa^{-1}L$ and hence L can be written as $L = aa^{-1}L \cup bb^{-1}L \cup (L \cap \{\varepsilon\})$. By assumption $aa^{-1}L$ and $bb^{-1}L$ have density. Because $aa^{-1}L$ and $bb^{-1}L$ are mutually disjoint, by the additivity of δ^*_A (Lemma 1-(4)) we have

$$\delta^*_A(L) = \delta^*_A(aa^{-1}L) + \delta^*_A(bb^{-1}L). \tag{3}$$

$K_n \subseteq L$ holds for each n hence we have $a^{-1}K_n \subseteq a^{-1}L$ and $\delta^*_A(a^{-1}K_n) \leq \delta^*_A(a^{-1}L)$. Because $(K_n)_n$ is a convergent sequence to L from inner, for any $\epsilon/2 > 0$ there exists δ such that $\delta^*_A(L) - \delta^*_A(K_n) < \epsilon/2$ holds for every $n > \delta$. Thus from Equality (3) we can deduce that

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

holds for every $n > \delta$. We know $\delta^*_A(cc^{-1}L') = \delta^*_A(c^{-1}L')/2$ for any $c \in \{a, b\}$ and L' by Lemma 1-(5), the above inequality can be transformed as

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

Hence we can conclude that $\delta^*_A(a^{-1}L) - \delta^*_A(a^{-1}K_n) < \epsilon$ for every $n > \delta$, *i.e.*, $(a^{-1}K_n)_n$ is a convergent sequence to $a^{-1}L$ from inner. We can show that $(a^{-1}M_n)_n$ converges to L from outer by the same way. □

Corollary 2. *Let $\mathcal{C} \subseteq \mathcal{D}$ be language classes where $\mathcal{C}$ is closed under Boolean operations and left-and-right quotients and every language in $\mathcal{D}$ has a density. Then $\mathcal{C}$-measurability in $\mathcal{D}$ is preserved under Boolean operations and left-and-right quotients.*

An application of Theorem 8 is a proof of the undecidability of REG-measurability for context-free languages, modulo the following conjecture.

Conjecture 1. If a context-free language L has a density, then its quotients $a^{-1}L$ and La^{-1} also have densities.

Theorem 9. *If Conjecture 1 is true, then it is undecidable whether a given context-free grammar generates* REG*-measurable language or not.*

Proof. The class CFL is closed under left-and-right quotients, hence by Theorem 8 the class $P = \{L \in \mathrm{CFL}_A \mid L \text{ is REG-measurable}\}_A$ is also closed under left-and-right quotients. It is clear that $\mathrm{REG} \subseteq P$ holds, and by Theorem 4-(1) there is REG-immeasurable context-free language M_2, *i.e.*, $P \subsetneq \mathrm{CFL}$. Because the universality problem for CFL is undecidable, the REG-measurability is also undecidable for CFL by the well-known Greibach's theorem [12]. □

We conclude this section by giving the following Carathéodory's condition characterisation of REG-measurability. The proof is almost same with one of Lebesgue measurability (*cf.* [2]), albeit that requires some density condition which is formal language theoretic.

Theorem 10. *Let $\mathcal{C}$ be a class of languages closed under Boolean operations and let $L \subseteq A^*$ be a language. If every language obtained by a finite Boolean combination of languages in $\mathcal{C} \cup \{L\}$ has a density, then L is $\mathcal{C}$-measurable if and only if the following Carathéodory's condition holds:*

$$\forall X \subseteq A^* \qquad \overline{\mu}_{\mathcal{C}}(X) = \overline{\mu}_{\mathcal{C}}(X \cap L) + \overline{\mu}_{\mathcal{C}}(X \cap \overline{L}). \tag{4}$$

Proof. If L satisfies the Carathéodory condition (4), then we obtain $\overline{\mu}_{\mathcal{C}}(A^*) = 1 = \overline{\mu}_{\mathcal{C}}(L) + \overline{\mu}_{\mathcal{C}}(\overline{L})$ when $X = A^*$, thus by Lemma 3, L is $\mathcal{C}$-measurable because $\mathcal{C}$ is closed under complementation by assumption.

Now we show the converse direction. Assume L is $\mathcal{C}$-measurable. For any language $X \subseteq A^*$ and for any $\epsilon > 0$, by the definition of $\overline{\mu}_{\mathcal{C}}$, there exists $K \in \mathcal{C}$ such that $X \subseteq K$ and $\delta^*_A(K) \leq \overline{\mu}_{\mathcal{C}}(X) + \epsilon$. Here L, K and $\overline{K}$ are all $\mathcal{C}$-measurable, and by assumption $K \cap L$ and $K \cap \overline{L}$ have densities. Hence, by Theorem 7, $K \cap L$ and $K \cap \overline{L}$ are $\mathcal{C}$-measurable. Because $K = (K \cap L) \cup (K \cap \overline{L})$ and $(K \cap L) \cap (K \cap \overline{L}) = \emptyset$,

$$\delta^*_A(K) = \delta^*_A(K \cap L) + \delta^*_A(K \cap \overline{L})$$

holds by the additivity of δ^*_A (Lemma 1-(4)). Hence we have

$$\begin{aligned} \overline{\mu}_{\mathcal{C}}(X) &\geq \delta^*_A(K) - \epsilon = \delta^*_A(K \cap L) + \delta^*_A(K \cap \overline{L}) - \epsilon \\ &\geq \overline{\mu}_{\mathcal{C}}(X \cap L) + \overline{\mu}_{\mathcal{C}}(X \cap \overline{L}) - \epsilon. \end{aligned}$$

Because $\epsilon > 0$ is taken arbitrarily, we can conclude that

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

holds. The reverse direction $\leq$ of the above inequality is directly obtained by the subadditivity of $\overline{\mu}_{\mathcal{C}}$ (Lemma 2-(3)). □

4 Carathéodory Extensions of Local Varieties

In this section, as a "miniature" of REG-measurability, we investigate $\mathcal{C}$-measurability for some subclass $\mathcal{C}$ of REG. The considered subclasses of regular languages here are so-called local varieties, which enjoy good closure properties and have rich algebraic structure. First we introduce some background materials from algebraic language theory.

4.1 Local varieties and an Eilenberg-type theorem

Due to the space limitation, we assume that the author has a basic knowledge of algebraic language theory (*e.g.*, syntactic monoids and morphism, *etc. cf.* [13, 14]). For a language L over A, we denote its syntactic monoid by $\mathsf{Synt}(L)$ and its syntactic morphism by $\eta_L : A^* \to \mathsf{Synt}(L)$. A monoid M is said to be *aperiodic*, if there is $k \geq 1$ such that $x^k = x^{k+1}$ for any $x \in M$. M is called *zero* if it contains zero element 0: $0 \cdot x = x \cdot 0 = 0$ for all $x \in M$. Further, a zero semigroup S is called *nilpotent* if there is $k \geq 1$ such that $x^k = 0$ for any $x \in S$. A non-empty subset $I \subseteq M$ is called ideal if $M \cdot I \cdot M \subseteq I$. An ideal I is said to be minimal if no proper subset of I is an ideal. It is well-known that any finite monoid has a unique minimal ideal (*cf.* [14]).

The main targets in the next subsection are classes of regular languages with some *good* closure properties as follows.

Definition 3 (*cf.* [15]). A family $\mathcal{C} \subseteq \mathrm{REG}_A$ of regular languages over A is called *local variety* if it is closed under Boolean operations and left-and-right quotients. A family V of finite monoids generated by A is called *local pseudovariety* if it is closed under quotients and subdirect products.

Theorem 11 (Eilenberg-type theorem for local varieties [16]). *For each A, there is a lattice isomorphism between the class of all local varieties and the class of all local pseudovarieties.*

This Eilenberg-type theorem roughly states that: if a class of languages is somewhat "robust" (*i.e.*, enjoys good closure properties), then it could be characterised by an algebraic way (at least there should exist the corresponding local pseudovariety), and vice versa. We now enumerate three examples of local varieties and corresponding local pseudovarieties (see Fig 1). A prominent example of a local variety is *star-free languages*. A language L is said to be star-free if it can be obtained by a finite combination of Boolean operations and concatenations of finite languages. The family SF_A of all star-free languages over A forms a local variety, and this class can be characterised in purely algebraic way as follows.

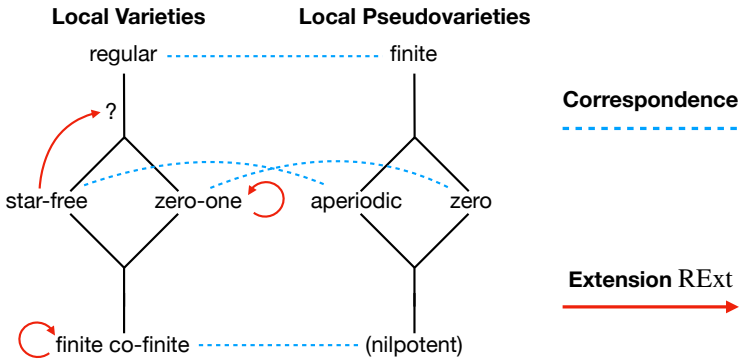

Fig. 1. Relation between local varieties, extensions and local pseudovarieties.

Theorem 12 (Schützenberger's theorem [17]).
The corresponding local pseudovariety of SF_A *is the class of aperiodic finite monoids generated by* A*. Namely,* $L \in \mathrm{SF}_A$ *if and only if* $\mathsf{Synt}(L)$ *is aperiodic.*

Next we introduce two additional examples of local varieties. One is the family FIN_A of all *finite and co-finite languages* and another one is the family ZO_A of all regular languages with *density either zero or one*. FIN_A and ZO_A form a local variety, respectively. In his Volume B [18], Eilenberg showed that the class of all finite nilpotent semigroups form a pseudovariety (of semigroups) and its corresponding +-variety of languages is exactly the class of all finite and co-finite languages. The corresponding local pseudovariety of ZO_A is the family of all finite zero monoids (*cf.* [6]).

4.2 Extension as a closure operator

In this subsection we mainly consider "extensions" of local varieties. All results are summarised in Fig. 1. First we introduce necessary notation.

Definition 4. For a family $\mathcal{C} \subseteq 2^{A^*}$ of languages over A, we define its *(Carathéodory) extension* as

$$\mathrm{Ext}_A(\mathcal{C}) = \{L \subseteq A^* \mid L \text{ is } \mathcal{C}\text{-measurable}\},$$

and define its *regular extension* as

$$\mathrm{REXt}_A(\mathcal{C}) = \mathrm{Ext}_A(\mathcal{C}) \cap \mathrm{REG}_A.$$

Observe that this extension operator is a closure as follows.

Theorem 13. Ext_A *is a closure operator, i.e., it satisfies the following three properties for each* $\mathcal{C}, \mathcal{D} \subseteq 2^{A^*}$*.*

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

monotone: $\mathcal{C} \subseteq \mathcal{D}$ *implies* $\mathrm{Ext}_A(\mathcal{C}) \subseteq \mathrm{Ext}_A(\mathcal{D})$.
idempotent: $\mathrm{Ext}_A(\mathrm{Ext}_A(\mathcal{C})) = \mathrm{Ext}_A(\mathcal{C})$.

Proof. The extensivity and monotonicity are clear from the definition. To show the idempotency, consider $L \in \mathrm{Ext}_A(\mathrm{Ext}_A(\mathcal{C}))$. In this case, there exists an infinite sequence $(K_n, M_n)_n$ of pairs of languages in $\mathrm{Ext}_A(\mathcal{C})$ that converges to L. For each n, K_n and M_n belong to $\mathrm{Ext}_A(\mathcal{C})$, thus there exist a sequence $(K_{(n,i)})_i$ of languages in $\mathcal{C}$ that converges to K_n from inner and a sequence $(M_{(n,i)})_i$ of languages in $\mathcal{C}$ that converges to M_n from outer. Then the infinite sequence $(K_{(n,n)}, M_{(n,n)})_n$ of pairs of languages in $\mathcal{C}$ converges to L,*i.e.*, $L \in \mathrm{Ext}_A(\mathcal{C})$. □

In the previous section, we showed that the $\mathcal{C}$-measurability is closed under Boolean operations and left-and-right quotients if $\mathcal{C}$ is closed under these operations and every language in $\mathcal{C}$ have a density (Corollary 2). Because every regular language have a density (Theorem 1), we have the following corollary.

Corollary 3. *For any local variety* $\mathcal{C} \subseteq \mathrm{REG}_A$ *over* A, $\mathrm{RExt}_A(\mathcal{C}) \supseteq \mathcal{C}$ *is also a local variety over* A, *i.e.,* RExt_A *is a closure operator over the class of all local varieties.*

Clearly, any FIN_A-measurable language is either finite or co-finite. A similar argument can be applied for ZO_A. Thus RExt_A does not properly extend these two local varieties.

Theorem 14. $\mathrm{RExt}_A(\mathrm{FIN}_A) = \mathrm{FIN}_A$ *and* $\mathrm{RExt}_A(\mathrm{ZO}_A) = \mathrm{ZO}_A$ *for each* A.

Furthermore, for a unary alphabet $A = \{a\}$, it is well-known that $\mathrm{SF}_A = \mathrm{ZO}_A = \mathrm{FIN}_A$, hence we have the following as a corollary.

Corollary 4. $\mathrm{RExt}_A(\mathrm{SF}_A) = \mathrm{SF}_A = \mathrm{ZO}_A = \mathrm{FIN}_A$ *for* $A = \{a\}$.

The situation is different for the case $\#(A) \geq 2$. As we explained in Remark 1, if $\#(A) \geq 2$, $\mathrm{RExt}_A(\mathrm{SF}_A)$ can contain *any* regular language over $B \subsetneq A$ hence $\mathrm{RExt}_A(\mathrm{SF}_A) \supsetneq \mathrm{SF}_A$ ($\mathrm{RExt}_A(\mathrm{SF}_A) \ni (aa)^* \notin \mathrm{SF}_A$, in particular). The next theorem says, however, $\mathrm{RExt}_A(\mathrm{SF}_A)$ can not contain some regular languages over A like $(AA)^*$.

Theorem 15. *If a star-free language* $L \in \mathrm{SF}_A$ *satisfies* $\delta^*_A(L) > 0$, *then* L *contains words of even and odd length.*

Proof. Consider the syntactic monoid $\mathsf{Synt}(L)$, the syntactic morphism $\eta_L : A^* \to \mathsf{Synt}(L)$ and the syntactic image $S = \eta_L(L)$ of L. Because L is regular, $\mathsf{Synt}(L)$ is finite. Hence it has a unique minimal ideal $K \subseteq \mathsf{Synt}(L)$. Let w_x be a word whose syntactic image $\eta_L(w_x)$ is x for each $x \in \mathsf{Synt}(L)$.

The assumption $\delta^*_A(L) > 0$ implies $S \cap K \neq \emptyset$, because $\delta^*_A(\eta^{-1}(\mathsf{Synt}(L) \setminus K)) = 0$ holds; for each $k \in K$ and $x, y \in A^*$, we have $\eta_L(xw_ky) = \eta_L(x) \cdot k \cdot \eta_L(y) \in K$ and hence $\eta^{-1}(\mathsf{Synt}(L) \setminus K) \cap A^* w_k A^* = \emptyset$ which implies $\eta^{-1}(\mathsf{Synt}(L) \setminus K)$ is null by infinite monkey theorem. Thus L is not null implies its syntactic image S contains at least one element of K, say, $t \in S \cap K$.

Clearly, $\delta_A^*(\eta^{-1}(K)) = 1$ holds and hence $\eta^{-1}(K)$ contains some word w_{odd} of odd length. Let $m_{\mathrm{odd}} = \eta_L(w_{\mathrm{odd}})$ be its syntactic image. By Schützenberger's theorem (Theorem 12), $\mathsf{Synt}(L)$ is aperiodic thus there is some $i \geq 1$ such that $m_{\mathrm{odd}}^i = m_{\mathrm{odd}}^{i+1}$. By the minimality of the ideal K, there exist $x, y \in \mathsf{Synt}(L)$ such that $x \cdot m_{\mathrm{odd}}^i \cdot y = t$ (if not, the ideal $\mathsf{Synt}(L) \cdot m_{\mathrm{odd}}^i \cdot \mathsf{Synt}(L)$ generated by m_{odd}^i does not contain t hence it should be a proper subset of K). Then two words $w_x w_{\mathrm{odd}}^i w_y$ and $w_x w_{\mathrm{odd}}^{i+1} w_y$ has the same syntactic image

$$\eta_L(w_x w_{\mathrm{odd}}^i w_y) = x \cdot m_{\mathrm{odd}}^i \cdot y = t = x \cdot m_{\mathrm{odd}}^{i+1} \cdot y = \eta_L(w_x w_{\mathrm{odd}}^{i+1} w_y),$$

thus both belong to L. Because the length of w_{odd} is odd, the lengths of these two words are different modulo 2. □

The above theorem tells us that any star-free subset of $(AA)^*$ is null and any star-free superset of $(AA)^*$ is co-null, thus we have the following corollary.

Corollary 5. *$(AA)^* \notin \mathrm{REXT}_A(\mathrm{SF}_A)$ for any A. In particular, $\underline{\mu}_{\mathrm{SF}_A}((AA)^*) = 0$ and $\overline{\mu}_{\mathrm{SF}_A}((AA)^*) = 1$. Further, $\mathrm{SF}_A \subsetneq \mathrm{REXT}_A(\mathrm{SF}_A) \subsetneq \mathrm{REG}_A$ if $\#(A) \geq 2$.*

We are not aware what the associated local pseudovariety of this new local variety $\mathrm{REXT}_A(\mathrm{SF}_A)$ yet, but, we can say that $\mathrm{REXT}_A(\mathrm{SF}_A)$ always contains all zero-one regular languages.

Theorem 16. *$\mathrm{REXT}_A(\mathrm{SF}_A) \supseteq \mathrm{ZO}_A$ for any A.*

Proof. The case $\#(A) = 1$ follows from Theorem 4. We show this for a general alphabet A. Let $L \in \mathrm{ZO}_A$ and we can assume $\delta_A^*(L) = 0$ without loss of generality. By Theorem 2, L is null implies there is some forbidden word w of L: $L \cap A^* w A^* = \emptyset$. Hence $L \subseteq \overline{A^* w A^*}$ holds and $\underline{\mu}_{\mathrm{SF}_A}(L) = \overline{\mu}_{\mathrm{SF}_A}(L) = 0$. □

5 Future Work and Open Problems

We have investigated general properties of $\mathcal{C}$-measurability, and examine how the extension operator REXT_A extends certain local varieties of regular languages. An immediate future work is to give an algebraic characterisation of $\mathrm{REXT}_A(\mathrm{SF}_A)$. We are also interested whether we can characterise the associated extension operator of local pseudovarieties of finite monoids $\mathrm{MEXT}_A(\mathsf{V}) = F(\mathrm{REXT}_A(F^{-1}(\mathsf{V})))$ in purely algebraic way, where F is the lattice isomorphism stated in Theorem 11. One of the ideal goals is to understand the class of REG-measurable context-free languages. However, it looks like a bit difficult since the theory of densities of context-free languages is not well developed yet (*e.g.*, Conjecture 1). Actually, we are not aware whether there is a context-free language that do not have a density ($L_\perp$ in Example 1-(3) is not context-free). More open problems related to REG-measurability and context-free languages were posed in [1].

Acknowledgements: This work was supported by JSPS KAKENHI Grant Number JP19K14582.

References

1. Sin'ya, R.: Asymptotic approximation by regular languages. In: Current Trends in Theory and Practice of Computer Science. (2021) 74–88
2. Tao, T.: An Introduction to Measure Theory. Graduate studies in mathematics. American Mathematical Society (2013)
3. Buck, R.C.: The measure theoretic approach to density. American Journal of Mathematics **68**(4) (1946) 560–580
4. Berstel, J., Perrin, D., Reutenauer, C.: Codes and Automata. Encyclopedia of Mathematics and its Applications. Cambridge University Press (2009)
5. Salomaa, A., Soittola, M.: Automata Theoretic Aspects of Formal Power Series. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1978)
6. Sin'ya, R.: An automata theoretic approach to the zero-one law for regular languages. In: Games, Automata, Logics and Formal Verification. (2015) 172–185
7. Flajolet, P.: Ambiguity and transcendence. In: Automata, Languages and Programming, Berlin, Heidelberg, Springer Berlin Heidelberg (1985) 179–188
8. Flajolet, P.: Analytic models and ambiguity of context-free languages. Theoretical Computer Science **49**(2) (1987) 283–309
9. Chomsky, N., Schützenberger, M.: The algebraic theory of context-free languages. In: Computer Programming and Formal Systems. Volume 35. (1963) 118–161
10. Kemp, R.: A note on the density of inherently ambiguous context-free languages. Acta Informatica **14**(3) (1980) 295–298
11. Dömösi, P., Horváth, S., Ito, M.: On the connection between formal languages and primitive words. In: First Session on Scientific Communication. (1991) 59–67
12. Greibach, S.A.: A note on undecidable properties of formal languages. Mathematical systems theory **2** (1968) 1–6
13. Lawson, M.V.: Finite Automata. Birkhäuser (2005)
14. Pin, J.E.: Mathematical foundations of automata theory (draft)
15. Adámek, J., Milius, S., Myers, R.S.R., Urbat, H.: Generalized eilenberg theorem i: Local varieties of languages. In: Foundations of Software Science and Computation Structures. (2014) 366–380
16. Gehrke, M., Grigorieff, S., Pin, J.: Duality and equational theory of regular languages. In: Automata, Languages and Programming. (2008) 246–257
17. Schützenberger, M.P.: On finite monoids having only trivial subgroups. Information and Control **8**(2) (1965) 190–194
18. Eilenberg, S., Tilson, B.: Automata, languages and machines. Volume B. Pure and applied mathematics. Academic Press, New-York, San Franciso, London (1976)