Asymptotic Approximation by Regular Languages

Ryoma Sin'ya
Akita University, Akita, Japan
RIKEN AIP, Japan
ryoma@math.akita-u.ac.jp

Abstract

This paper investigates a new property of formal languages called REG-measurability where REG is the class of regular languages. Intuitively, a language L is REG-measurable if there exists an infinite sequence of regular languages that "converges" to L. A language without REG-measurability has a complex shape in some sense so that it can not be (asymptotically) approximated by regular languages. We show that several context-free languages are REG-measurable (including languages with transcendental generating function and transcendental density, in particular), while a certain simple deterministic context-free language and the set of primitive words are REG-immeasurable in a strong sense.

2012 ACM Subject Classification Theory of computation → Formal languages and automata theory

Keywords and phrases Automata, context-free languages, density, primitive words

Digital Object Identifier 10.4230/LIPIcs.CVIT.2016.23

Funding *Ryoma Sin'ya*: JSPS KAKENHI Grant Number JP19K14582

1 Introduction

Approximating a complex object by more simple objects is a major concept in both computer science and mathematics. In the theory of formal languages, various types of approximations have been investigated (*e.g.*, [14, 15, 10, 7, 5, 8]). For example, Kappes and Kintala [14] introduced *convergent-reliability* and *slender-reliability* which measure how a given deterministic automaton $\mathcal{A}$ nicely approximates a given language L over an alphabet A. Formally $\mathcal{A}$ is said to accept L convergent-reliability if the ratio of the number of *in*correctly accepted/rejected words of length n

$$\#((L(\mathcal{A}) \triangle L) \cap A^n) / \#(A^n)$$

tends to 0 if n tends to infinity, and is said to accept L slender-reliability if the number of incorrectly accepted/rejected words of length n is always bounded above by some constant c: *i.e.*, $\#((L(\mathcal{A}) \triangle L) \cap A^n) \leq c$ for any n. Here $L(\mathcal{A})$ denotes the language accepted by $\mathcal{A}$, $\#(S)$ denotes the cardinality of the set S, $\overline{L}$ denotes the complement of L and $\triangle$ denotes the symmetric difference. A slightly modified version of approximation is *bounded-ϵ-approximation* which was introduced by Eisman and Ravikumar. They say that two languages L_1 and L_2 provide a bounded-ϵ-approximation of language L if $L_1 \subseteq L \subseteq L_2$ holds and the ratio of their length-n difference satisfies

$$\#((L_2 \setminus L_1) \cap A^n) / \#(A^n) \leq \epsilon$$

for every sufficiently large $n \in \mathbb{N}$. Perhaps surprisingly, they showed that no pair of regular languages can provide a bounded-ϵ-approximation of the language $\{w \in \{a, b\}^* \mid w \text{ has more } a\text{'s than } b\text{'s}\}$ for any $0 \leq \epsilon < 1$ [10]. This result is a very strong *in*approximable (by regular languages) example of certain non-regular languages. Also, there is a different framework of approximation so-called *minimal-cover* [5, 8].

42nd Conference on Very Important Topics (CVIT 2016).
Editors: John Q. Open and Joan R. Access; Article No. 23; pp. 23:1–23:16
Leibniz International Proceedings in Informatics
LIPICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

A model of approximation introduced in this paper is rather close to the work of Eisman and Ravikumar [10]. Instead of approximating by a *single* regular language, we consider an approximation of some non-regular language L by an *infinite sequence* of regular languages that "converges" to L. Intuitively, we say that L is REG-*measurable* if there exists an infinite sequence of pairs of regular languages $(K_n, M_n)_{n\in\mathbb{N}}$ such that $K_n \subseteq L \subseteq M_n$ holds for all n and the "size" of the difference $M_n \setminus K_n$ tends to 0 if n tends to infinity. The formal definition of "size" is formally described in the next section: we use a notion called *density (of languages)* for measuring the "size" of a language.

Although we used the term "approximation" in the title and there are various research on this topic in formal language theory, our work is strongly influenced by the work of Buck [4] which investigates, as the title said, *the measure theoretic approach to density.* In [4] the concept of *measure density* μ of subsets of natural numbers $\mathbb{N}$ was introduced. Roughly speaking, Buck considered an arithmetic progression $X = \{cn + d \mid n \in \mathbb{N}\}$ (where $c, d \in \mathbb{N}$, c can be zero) as a "basic set" whose *natural density* as $\delta(X) = 1/c$ if $c \neq 0$ and $\delta(X) = 0$ otherwise, then defined the *outer measure density* $\mu^*(S)$ of any subset $S \subseteq \mathbb{N}$ as

$$\mu^*(S) = \inf\Big\{\sum_i \delta(X_i) \mid S \subseteq X \text{ and } X \text{ is a finite union of disjoint arithmetic progressions } X_1, \dots, X_k\Big\}.$$

Then the *measure density* $\mu(S) = \mu^*(S)$ was introduced for the sets satisfying the condition

$$\mu^*(S) + \mu^*(\overline{S}) = 1 \tag{1}$$

where $\overline{S} = \mathbb{N} \setminus S$. Technically speaking, the class $\mathcal{D}_\mu$ of all subsets of natural numbers satisfying Condition (1) is the *Carathéodory extension* of the class

$$\mathcal{D}_0 \overset{\text{def}}{=} \{X \subseteq \mathbb{N} \mid X \text{ is a finite union of arithmetic progressions }\},$$

see Section 2 of [4] for more details. Notice that here we regard a singleton $\{d\}$ as an arithmetic progression (the case $c = 0$ for $\{cn + d \mid n \in \mathbb{N}\}$), any finite set belongs to $\mathcal{D}_0$. Buck investigated several properties of μ and $\mathcal{D}_\mu$, and showed that $\mathcal{D}_\mu$ *properly* contains $\mathcal{D}_0$.

In the setting of formal languages, it is very natural to consider the class REG of regular languages as "basic sets" since it has various types of representation, good closure properties and rich decidable properties. Moreover, if we consider regular languages REG_A over a unary alphabet $A = \{a\}$, then REG_A is isomorphic to the class $\mathcal{D}_0$; it is well known that the Parikh image $\{|w| \mid w \in L\} \subseteq \mathbb{N}$ (where $|w|$ denotes the length of w) of every regular language L in REG_A is semilinear and hence it is just a finite union of arithmetic progressions. From this observation, investigating the densities of regular languages and its measure densities (*i.e.*, REG-measurability) for non-regular languages can be naturally considered as an adaptation of Buck's study [4] for formal language theory.

Our contribution

In this paper we investigate REG-measurability ($\simeq$ asymptotic approximability by regular languages) of non-regular, mainly context-free languages. The main results consist of three kinds. We show that: (1) several context-free languages (including languages with *transcendental generating function* and *transcendental density*) are REG-measurable [Theorem 23–30]. (2) there are "very large/very small" (deterministic) context-free languages that are REG-immeasurable in a strong sense [Theorem 36]. (3) the set of *primitive words* is "very large" and REG-immeasurable in a strong sense [Theorem 37–38]. Open problems and some possibility of an application of the notion of measurability to classifying formal languages will be stated in Section 6.

The paper is organised as follows. Section 2 provides mathematical background of densities of formal languages. The formal definition of REG-approximability and REG-measurability are introduced in Section 3. The scenario of Section 3 mostly follows one of the measure density introduced by Buck [4] which was described above. In Section 4, we will give several examples of REG-inapproximable but REG-measurable context-free languages. These examples include, perhaps somewhat surprisingly, a language with a *transcendental density* which have been considered as a very complex context-free language from a combinatorial viewpoint. In Section 5, we consider the set of so-called *primitive words* and its REG-measurability. Section 6 ends this paper with concluding remarks, some future work and open problems. We assume that the reader has a basic knowledge of formal language theory.

2 Densities of Formal Languages

For a set S, we write $\#(S)$ for the cardinality of S. The set of natural numbers including 0 is denoted by $\mathbb{N}$. For an alphabet A, we denote the set of all words (resp. all non-empty words) over A by A^* (resp. A^+). We write ε for the empty word and write A^n (resp. $A^{<n}$) for the set of all words of length n (resp. less than n). For a language L, we write $\mathtt{Alph}(L)$ for the set of all letters appeared in L. For 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 *factor* of a word w if $w = xvy$ for some $x, y \in A^*$, further said to be a *prefix* of w if $x = \varepsilon$. For a language $L \subseteq A^*$, we denote by $\overline{L} = A^* \setminus L$ the complement of L.

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 $\mathrm{REG}, \mathrm{DetCFL}, \mathrm{UnCFL}$ and CFL the class of regular languages, deterministic context-free languages, unambiguous context-free languages and context-free languages, respectively. A language L is said to be $\mathcal{C}$*-immune* if L is infinite and no infinite subset of L belongs to $\mathcal{C}$.

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

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

if the limit exists, otherwise we write $\delta_A(L) = \perp$ and say that L does not have a natural density. The *density* $\delta^*_A(L)$ of L is defined as

$$\delta^*_A(L) \stackrel{\mathtt{def}}{=} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \frac{\#\left(L \cap A^k\right)}{\#(A^k)}$$

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

▶ Remark 2. Notice that if L has a natural density (*i.e.*, $\delta_A(L) \neq \perp$), then it also has a density and $\delta^*_A(L) = \delta_A(L)$ holds. But the converse is not true in general, *e.g.*, the case $L = (AA)^*$ (see Example 4 below).

The following observation is basic.

▷ Claim 3. 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 $K \subseteq L$.
3. $\delta^*_A(\overline{K}) = 1 - \alpha$.
4. $\delta^*_A(K \cup L) \leq \alpha + \beta$ if $\delta^*_A(K \cup L) \neq \bot$.
5. $\delta^*_A(K \cup L) = \alpha + \beta$ if $K \cap L = \emptyset$.

For more properties of δ^*_A, see Chapter 13 of [3].

▶ **Example 4.** Here we enumerate a few examples of densities of languages.
- The set of all words A^* clearly satisfies $\delta_A(A^*) = 1$, and its complement $\emptyset$ satisfies $\delta_A(\emptyset) = 0$. It is also clear that every finite language is null.
- For the set $\{a\}A^*$ of all words starting with $a \in A$, we have $\#(\{a\}A^* \cap A^n) / \#(A^n) = \#(aA^{n-1}) / \#(A^n) = 1/\#(A)$. Hence $\delta_A(\{a\}A^*) = 1/\#(A)$.
- Consider $(AA)^*$ the set of all words with even length. Because

$$\frac{\#((AA)^* \cap A^n)}{\#(A^n)} = \begin{cases} 1 & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

  holds, its limit does not exist and thus $(AA)^*$ does not have a natural density $\delta_A((AA)^*) = \bot$. However, it has a density $\delta^*_A((AA)^*) = 1/2$.
- The semi-Dyck language

$$\mathsf{D} \stackrel{\text{def}}{=} \{w \in \{a,b\}^* \mid |w|_a = |w|_b \text{ and } |u|_a \geq |u|_b \text{ for every prefix } u \text{ of } w\}$$

  is non-regular but context-free. It is well known that the number of words in D of length $2n$ is equal to the n-th Catalan number whose asymptotic approximation is $\Theta(4^n/n^{3/2})$. Thus

$$\frac{\#(\mathsf{D} \cap A^n)}{\#(A^n)} = \begin{cases} \Theta(1/(n/2)^{3/2}) & \text{if } n \text{ is even,} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

  and we have $\delta_A(\mathsf{D}) = 0$, *i.e.*, D is null.

Example 4 shows us that, for some regular language L, its natural density is either zero or one, for some, like $L = \{a\}A^*$ (for $\#(A) \geq 2$), $\delta_A(L)$ could be a real number strictly between zero and one, and for some, like $L = (AA)^*$, a natural density may not even exist. However, the following theorem tells us that all regular languages *do* have densities.

▶ **Theorem 5** (*cf.* Theorem III.6.1 of [20]). *Let $L \subseteq A^*$ be a regular language. Then there is a positive integer c such that for all natural numbers $d < c$, the following limit exists*

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

and it is always rational, i.e., the sequence $(\#(L \cap A^n) / \#(A^n))_{n \in \mathbb{N}}$ has only finitely many accumulation points and these are rational and periodic.

▶ **Corollary 6.** *Every regular language has a density and it is rational.*

▶ **Corollary 7.** *For any regular language $L \subseteq A^*$, $\delta_A(L) = 0$ if and only if $\delta^*_A(L) = 0$.*

Furthermore, for *unambiguous* context-free languages, the following holds.

▶ **Theorem 8** (Berstel [2]). *For any unambiguous context-free language L over A, its density $\delta^*_A(L)$, if it exists (i.e., $\delta^*_A(L) \neq \bot$), is always algebraic.*

In the next section we will introduce a language with a transcendental density, which should be inherently ambiguous due to Theorem 8.

We conclude the section by introducing the notion called *dense*: a property about some topological "largeness" of a language (*cf.* Chapter 2.5 of [3]).

▶ **Definition 9.** A language $L \subseteq A^*$ is said to be *dense* if the set of all factors of L is equal to A^*. We say that a word $w \in A^*$ is a *forbidden word* (resp. *forbidden prefix*) of L if $L \cap A^*wA^* = \emptyset$ (resp. $L \cap wA^* = \emptyset$).

Observe that $L \subseteq A^*$ is dense if and only if no word is a forbidden word of L. The next theorem ties two different notions of "largeness" of languages in the regular case.

▶ **Theorem 10** (S. [21]). *A regular language is non-null if and only if it is dense.*

The "only if"-part of Theorem 10 is nothing but the well-known so-called *infinite monkey theorem* (which states that L is not dense implies L is null), and this part is true for any (non-regular) languages. But we stress that "if"-part is *not true* beyond regular languages; for example the semi-Dyck language D is null *but dense* (which will be described in Proposition 12). We denote by REG^+ the family of non-null regular languages, which is equivalent to the family of regular languages with positive densities thanks to Corollary 6.

3 Approximability and Measurability

Although we will mainly consider REG-measurability of non-regular languages in this paper, here we define two notions approximability and measurability in general setting, with few concrete examples.

▶ **Definition 11.** Let $\mathcal{C}, \mathcal{D}$ be class of languages. A language L is said to be $(\mathcal{C}, \epsilon)$-*lower-approximable* if there exists $K \in \mathcal{C}$ such that $K \subseteq L$ and $\delta^*_{\mathtt{Alph}(L)}(L \setminus K) \leq \epsilon$. A language L is said to be $(\mathcal{C}, \epsilon)$-*upper-approximable* if there exists $M \in \mathcal{C}$ such that $L \subseteq M$ and $\delta^*_{\mathtt{Alph}(M)}(M \setminus L) \leq \epsilon$. A language L is said to be $\mathcal{C}$-*approximable* ($\mathcal{C}$-*approx.* for short) if L is both $(\mathcal{C}, 0)$-lower and $(\mathcal{C}, 0)$-upper-approximable. $\mathcal{D}$ is said to be $\mathcal{C}$-approx. if every language in $\mathcal{D}$ is $\mathcal{C}$-approx.

The following proposition gives a simple REG-inaproximable example.

▶ **Proposition 12.** *The semi-Dyck language* D *is* REG-*inapprox.*

Proof. We already mentioned that D is null in Example 4, and thus D is $(\mathrm{REG}, 0)$-lower-approx by $\emptyset \subseteq \mathsf{D}$. One can easily observe that D has no forbidden word: since for any $w \in A^*$ there exists a pair of natural numbers $(n, m) \in \mathbb{N}^2$ such that $a^n w b^m \in \mathsf{D}$. Hence if a regular language L satisfies $\mathsf{D} \subseteq L$, L has no forbidden word, too, and thus L is non-null by Theorem 10. Thus by Claim 3, $\delta^*_A(L \setminus \mathsf{D}) = \delta^*_A(L) - \delta^*_A(\mathsf{D}) = \delta^*_A(L) > 0$, which means that D can not $(\mathrm{REG}, 0)$-upper-approx. ◀

The proof of Proposition 12 only depends on the non-existence of forbidden words, hence we can apply the same proof to the next theorem.

▶ **Theorem 13.** *Any null language having no forbidden word is* $(\mathrm{REG}, 0)$-*upper-inapprox.*

Because D is deterministic context-free, in our term we have:

▶ **Corollary 14.** DetCFL *is* REG-*inapprox.*

Furthermore, by the combination of Theorem 8 and the next theorem, we will know that there exists a context-free language which can not be approximated by any unambiguous context-free language.

▶ **Theorem 15** (Kemp [16]). *Let $A = \{a, b, c\}$. Define*

$$S_1 \stackrel{\text{def}}{=} \{a\}\{b^i a^i \mid i \geq 1\}^* \qquad S_2 \stackrel{\text{def}}{=} \{a^i b^{2i} \mid i \geq 1\}^*\{a\}^+,$$

and

$$L_1 \stackrel{\text{def}}{=} S_1\{c\}A^* \qquad L_2 \stackrel{\text{def}}{=} S_2\{c\}A^*.$$

Then $\mathsf{K} \stackrel{\text{def}}{=} L_1 \cup L_2$ *is a context-free language with a transcendental natural density* $\delta_A(\mathsf{K})$*.*

▶ **Corollary 16.** CFL *is* UnCFL*-inapprox.*

We then introduce the notion of $\mathcal{C}$-measurability which is a formal language theoretic analogue of Buck's measure density [4].

▶ **Definition 17.** Let $\mathcal{C}, \mathcal{D}$ be classes of languages. For a language L, we define its $\mathcal{C}$*-lower-density* as

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

and its $\mathcal{C}$*-upper-density* as

$$\overline{\mu}_{\mathcal{C}}(L) \stackrel{\text{def}}{=} \inf\{\delta^*_A(K) \mid A = \mathtt{Alph}(L), L \subseteq K, K \in \mathcal{C}_A, \delta^*_A(K) \neq \bot\}.$$

A language L is said to be $\mathcal{C}$*-measurable* if $\overline{\mu}_{\mathcal{C}}(L) = \underline{\mu}_{\mathcal{C}}(L)$ holds, and we simply write $\overline{\mu}_{\mathcal{C}}(L)$ as $\mu_{\mathcal{C}}(L)$. $\mathcal{D}$ is said to be $\mathcal{C}$-measurable if every language in $\mathcal{D}$ is $\mathcal{C}$-measurable.

▶ **Definition 18.** We call $\overline{\mu}_{\mathcal{C}}(L) - \underline{\mu}_{\mathcal{C}}(L)$ the $\mathcal{C}$*-gap* of a language L. We say that a language L *has full* $\mathcal{C}$*-gap* if its $\mathcal{C}$-gap equals to 1, *i.e.*, $\overline{\mu}_{\mathcal{C}}(L) - \underline{\mu}_{\mathcal{C}}(L) = 1$.

In the next section, we describe several examples of both REG-measurable and REG-immeasurable languages. The REG-gap could be a good measure how much a given language has a complex shape from the viewpoint of regular languages.

The following lemmata are basic.

▶ **Lemma 19.** *Let K, L be two languages.*
1. $\overline{\mu}_{\mathcal{C}}(K) \leq \overline{\mu}_{\mathcal{C}}(L)$ *if* $K \subseteq L$*.*
2. $\overline{\mu}_{\mathcal{C}}(K \cup L) \leq \overline{\mu}_{\mathcal{C}}(K) + \overline{\mu}_{\mathcal{C}}(L)$ *if* $\mathcal{C}$ *is closed under union.*
3. $\overline{\mu}_{\mathcal{C}}(K) = \delta^*_A(K)$ *if* $K \in \mathcal{C}$ *and* $\delta^*_A(K) \neq \bot$*.*

▶ **Lemma 20.** *Let $\mathcal{C}$ be a language class such that $\mathcal{C}$ is closed under complement and every language in $\mathcal{C}$ has a density. A language $L \subseteq A^*$ is $\mathcal{C}$-measurable if and only if*

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

Proof. Let L be a language and $A = \mathtt{Alph}(L)$. By definition, L satisfies Condition (2) if and only if

$$\inf\{\delta^*_A(K) \mid L \subseteq K, K \in \mathcal{C}\} = 1 - \inf\{\delta^*_A(K) \mid \overline{L} \subseteq K, K \in \mathcal{C}\} \tag{3}$$

holds. On the other hand, L is measurable if and only if

$$\inf\{\delta^*_A(K) \mid L \subseteq K, K \in \mathcal{C}\} = \sup\{\delta^*_A(K) \mid K \subseteq L, K \in \mathcal{C}\}. \tag{4}$$

For any language $K \in \mathcal{C}_A$ such that $K \subseteq L$ and $\delta^*_A(K) \neq \bot$, its complement $\overline{K}$ satisfies $\overline{L} \subseteq \overline{K}$ and $\delta^*_A(\overline{K}) = 1 - \delta^*_A(K)$. This means that if $\mathcal{C}_A$ is closed under complement then $\sup\{\delta^*_A(K) \mid K \subseteq L, K \in \mathcal{C}_A\} = 1 - \inf\{\delta^*_A(K) \mid \overline{L} \subseteq K, K \in \mathcal{C}_A\}$, holds, which immediately implies the equivalence of Condition (3) and Condition (4). ◀

4 REG-measurability on Context-free Languages

In this section we examine REG-measurability of several types of context-free languages. The first type of languages (Section 4.1) is null context-free languages. Although some null language can have a full REG-gap as stated in the next theorem, we will show that typical null context-free languages are REG-measurable.

▶ **Theorem 21.** *There is a recursive language L which is null but $\overline{\mu}_{\mathrm{REG}}(L) = 1$.*

Proof. Let A be an alphabet with $\#(A) \geq 2$ and let $(\mathcal{A}_i)_{i \in \mathbb{N}}$ be an enumeration of automata over A such that $\mathrm{REG}_A = \{L(\mathcal{A}_i) \mid i \in \mathbb{N}\}$; we can take such enumeration by enumerating some binary representation of automata via shortlex order $<_{\mathrm{lex}}$. We will construct a null language L such that $\overline{\mu}_{\mathrm{REG}}(L) = 1$, in particular, L intersects with every regular infinite language.

Consider the following program P which takes an input word w:

Step 1 set $i = 0$ and $\ell = 0$.
Step 2 check $L(\mathcal{A}_i)$ is infinite or not.
Step 3 if $L(\mathcal{A}_i)$ is finite, then set $i = i + 1$ and go back to Step 2.
Step 4 otherwise, pick u such that u is the smallest (with respect to $<_{\mathrm{lex}}$) word satisfying $|u| > \ell$ and $u \in L$ (such u surely exists since $L(\mathcal{A}_i)$ is infinite).
Step 5 if $w = u$ then P accepts w and halts.
Step 6 if $w <_{\mathrm{lex}} u$ then P rejects w and halts.
Step 7 if $u <_{\mathrm{lex}} w$ then set $\ell = |u|$, $i = i + 1$ and go back to Step 2.

One can easily observe that all Steps are effective and P ultimately halts for any input word w because the length of the word u in Step 4 is strictly increasing until $u = w$ or $w <_{\mathrm{lex}} u$. Thus the language $L \overset{\text{def}}{=} \{w \in A^* \mid P \text{ accepts } w\}$ is recursive, (1) $L \cap R \neq \emptyset$ for any regular infinite language because by Step (4–5) P accepts some word $w \in R$, and (2) $\delta_A(L) = 0$; by Step (5–6) and the length of u is strictly increasing, P rejects every word in A^n except for one single word u, for each n. Thus $\delta_A(L) = 0$ and $\overline{\mu}_{\mathrm{REG}}(L) = 1$. ◀

The second type of languages (Section 4.2) is inherently ambiguous languages and the third type of languages (Section 4.3) includes Kemp's language K whose density is transcendental. The last type of languages (Section 4.4) is languages with full REG-gap, *i.e.*, strongly REG-immeasurable languages.

4.1 Null Context-free Languages

First we consider the following language with constraints on the number of occurrences of letters, which is a very typical example of a non-regular but context-free language.

▶ **Definition 22.** For an alphabet A and letters $a, b \in A$ such that $a \neq b$, we define

$$L_A(a, b) \overset{\text{def}}{=} \{w \in A^* \mid |w|_a = |w|_b\}.$$

▶ **Theorem 23.** *$L_A(a, b)$ is REG-measurable where $A = \{a, b\}$.*

Proof. It is enough to show that the complement $L = \overline{L(a, b)}$ satisfies $\underline{\mu}_{\mathrm{REG}}(L) = 1$. For each $k \geq 1$, we define

$$L_k \overset{\text{def}}{=} \{w \in A^* \mid |w|_a \neq |w|_b \mod k\}.$$

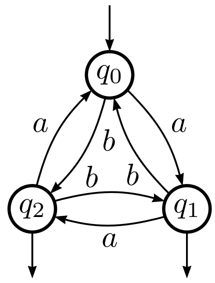

Figure 1 The deterministic automaton $\mathcal{A}_3$ in the Proof of Theorem 23. Here, the state q_0 having unlabelled incoming arrow is initial and the states q_1, q_2 having unlabelled outgoing arrow are final.

Clearly, $L_k \subseteq L$ holds. Each L_k is recognised by a k-states deterministic automaton

$$\mathcal{A}_k = (Q_k = \{q_0, \ldots, q_{k-1}\}, \Delta_k : Q_k \times A \to Q_k, q_0, Q_k \setminus \{q_0\})$$

where

$$\Delta_k(q_i, a) = q_{i+1 \bmod k} \qquad \Delta_k(q_i, b) = q_{i-1 \bmod k} \quad (\text{ for each } i \in \{0, \ldots, k-1\}),$$

q_0 is the initial state, and any other state $q \in Q_k \setminus \{q_0\}$ is a final state (the case $k = 3$ is depicted in Fig 1). The adjacency matrix of $\mathcal{A}_k$ is

$$M_k = \begin{bmatrix} 0 & 1 & 0 & \cdots & \cdots & 1 \\ 1 & 0 & 1 & \ddots & & \vdots \\ 0 & 1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 & 0 \\ \vdots & & \ddots & 1 & 0 & 1 \\ 1 & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix} = E_k + E_k^{k-1} \text{ where } E_k = \begin{bmatrix} 0 & 0 & 0 & \cdots & \cdots & 1 \\ 1 & 0 & 0 & \ddots & & \vdots \\ 0 & 1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & 0 \\ \vdots & & \ddots & 1 & 0 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix}.$$

M_k is a special case of *circulant matrices*. A k-dimensional circulant matrix C_k is a matrix that can be represented by a polynomial of E_k:

$$C_k = p(E_k) = \sum_{n=0}^{k-1} c_n E_k^n$$

and it is well known that C_k can be diagonalised as, for a k-th root of unity $\xi_k = e^{-\frac{2\pi i}{k}}$ (where i is the imaginary unit),

$$\frac{1}{\sqrt{k}} F_k^H \cdot C_k \cdot \frac{1}{\sqrt{k}} F_k = \operatorname{diag}(p(1), p(\xi_k^{-1}), p(\xi_k^{-2}), \ldots, p(\xi_k^{-(k-1)}))$$

where $F_k = (f_{n,m})$ with $f_{n,m} = \xi_k^{(n-1)(m-1)}$ (for $1 \leq n, m \leq k$) is the k-dimensional *Fourier matrix*, F_k^H is its Hermitian transpose and $\operatorname{diag}(\lambda_1, \cdots, \lambda_k)$ is the diagonal matrix whose n-th diagonal element is λ_n (for $1 \leq n \leq k$) (*cf.* Section 5.2.1 of [17]). Hence, in the case of $M_k = p_{\mathcal{A}_k}(E_k) = E_k + E_k^{k-1}$, we have

$$\frac{1}{\sqrt{k}} F_k^H \cdot M_k \cdot \frac{1}{\sqrt{k}} F_k = \operatorname{diag}(2, \xi_k^{-1} + \xi_k, \xi_k^{-2} + \xi_k^2, \ldots, \xi_k^{-(k-1)} + \xi_k^{k-1}) \tag{5}$$

because, for any $n \geq 0$, $p_{\mathcal{A}_k}(\xi_k^{-n}) = \xi_k^{-n} + \xi_k^{-n(k-1)} = \xi_k^{-n} + \xi_k^{n}$ holds.

Let $\Lambda_k = \mathrm{diag}(2, \xi_k^{-1} + \xi_k, \xi_k^{-2} + \xi_k^2, \ldots, \xi_k^{-(k-1)} + \xi_k^{k-1})$. Because $\mathcal{A}_k$ is deterministic and the final states are all but q_0, the number of words of length n in L_k is exactly the number of paths from q_0 to any other state in $\mathcal{A}_k$. For the k-dimensional vectors $\boldsymbol{e} = (1, 0, 0, \ldots, 0)$ and $\mathbf{1} = (1, 1, 1, \ldots, 1)$, from Equation (5) we have

$$\begin{aligned}
\#(L_k \cap A^n) &= \boldsymbol{e} \cdot M_k^n \cdot (\mathbf{1} - \boldsymbol{e})^T \\
&= \frac{1}{k} \boldsymbol{e} \cdot F_k \cdot \Lambda_k^n \cdot F_k^H (\mathbf{1} - \boldsymbol{e})^T \\
&= \frac{1}{k} \mathbf{1} \cdot \Lambda_k^n \cdot \left(k-1, \sum_{j=1}^{k-1} \xi_k^{-j}, \sum_{j=1}^{k-1} \xi_k^{-2j}, \ldots, \sum_{j=1}^{-(k-1)} \xi_k^{-(k-1)j} \right)^T \\
&= \frac{1}{k} \left(2^n (k-1) + (\xi_k^{-1} + \xi_k)^n \sum_{j=1}^{k-1} \xi_k^{-j} + \cdots + (\xi_k^{-(k-1)} + \xi_k^{k-1})^n \sum_{j=1}^{k-1} \xi_k^{-(k-1)j} \right). \qquad (6)
\end{aligned}$$

If k is odd $k = 2m+1$, then for any $1 \leq j \leq k-1$, $\xi_k^{-j} + \xi_k^j$ is a real number whose absolute value is strictly smaller than 2; because ξ_k^{-j} is the complex conjugate of ξ_k^j and hence $|\xi_k^{-j} + \xi_k^j| = |2\mathrm{Re}(\xi_k^j)| < 2$ for odd k. Hence from Equation (6) we can deduce that

$$\#(L_k \cap A^n) = \frac{k-1}{k} 2^n + o(2^n)$$

where $o(2^n)$ means some function such that $\lim_{n\to\infty} o(2^n)/2^n = 0$. Thus we have $\delta_A(L_k) = \frac{k-1}{k}$ for odd $k = 2m+1$, which tends to 1 if k tends infinity, *i.e.*, $\mu_{\mathrm{REG}}(L) = 1$. This completes the proof. ◀

By Theorem 23, it is also true that any subset of $L_{\{a,b\}}(a, b)$ is REG-measurable. In particular, we have:

▶ **Corollary 24.** *The semi-Dyck language* $\mathsf{D} \subseteq L_{\{a,b\}}(a, b)$ *is* REG*-measurable.*

The next example is the set of all palindromes.

▶ **Theorem 25.** $\mathsf{P}_A \overset{\mathsf{def}}{=} \{w \in A^* \mid w = \mathrm{rev}(w)\}$ *is* REG*-measurable.*

Proof. Because the case $\#(A) = 1$ is trivial ($\mathsf{P}_A = A^*$), we assume that $\#(A) \geq 2$. It is enough to show that the complement $\overline{\mathsf{P}_A}$ is REG-measurable.

For each $k \geq 1$, we define

$$L_k \overset{\mathsf{def}}{=} \{w_1 A^* w_2 \mid w_1, w_2 \in A^k, w_1 \neq \mathrm{rev}(w_2)\}.$$

One can easily observe that $L_k \subseteq \overline{\mathsf{P}_A}$ for each $k \geq 1$. Moreover, for any $n > 2k$, the number of words in L_k of length n is

$$\#(L_k \cap A^n) = \#(A)^k \cdot \#(A)^{n-2k} \cdot (\#(A)^k - 1) = \#(A)^n - \#(A)^{n-k}.$$

From this we can conclude that $\delta_A(L_k) = 1 - \#(A)^{-k}$ and it tends to 1 if k tends to infinity. Thus we have $\mu_{\mathrm{REG}}(\overline{\mathsf{P}_A}) = 1$. ◀

4.2 Some Inherently Ambiguous Languages

There are REG-measurable inherently ambiguous context-free languages. Since every *bounded language* $L \subseteq w_1^* \cdots w_k^*$ is trivially REG-measurable ($\mu_{\mathrm{REG}}(L) = 0$), a typical example of an inherently ambiguous context-free language $\{a^i b^j c^k \mid i = j \text{ or } i = k\}$ is REG-measurable.

Some more complex examples of inherently ambiguous languages are the following languages with constraints on the number of occurrences of letters investigated by Flajolet [12]:

$$\mathsf{O}_3 \stackrel{\text{def}}{=} \{w \in \{a,b,c\}^* \mid |w|_a = |w|_b \text{ or } |w|_a = |w|_c\},$$

$$\mathsf{O}_4 \stackrel{\text{def}}{=} \{w \in \{x,\bar{x},y,\bar{y}\}^* \mid |w|_x = |w|_{\bar{x}} \text{ or } |w|_y = |w|_{\bar{y}}\}.$$

▶ **Theorem 26.** O_3 *and* O_4 *are* REG*-measurable.*

Proof. Let $A = \{a,b,c\}$. For the case O_3, in a very similar way to Theorem 23, we can construct a sequence of automata $(\mathcal{A}_k^{ab})_{k\in\mathbb{N}}$ such that each automaton $\mathcal{A}_k^{ab}$ satisfies $L(\mathcal{A}_k^{ab}) \subseteq \overline{L_A(a,b)}$ and its adjacency matrix is of the form

$$M_k^{ab} = M_k + I_k = \begin{bmatrix} 1 & 1 & 0 & \cdots & \cdots & 1 \\ 1 & 1 & 1 & \ddots & & \vdots \\ 0 & 1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 & 0 \\ \vdots & & \ddots & 1 & 1 & 1 \\ 1 & \cdots & \cdots & 0 & 1 & 1 \end{bmatrix}$$

where M_k is the adjacency matrix stated in Theorem 23 and I_k is the k-dimensional identity matrix. The automaton $\mathcal{A}_k^{ab}$ is obtained by just adding self-loop labeled by c for each state $q \in Q_k$ of $\mathcal{A}_k$ in Theorem 23. This sequence of automata ensures that the language $L_A(a,b)$ is REG-measurable ($\overline{\mu}_{\mathrm{REG}}(L_A(a,b)) = 0$, in particular). The same argument is applicable to the language $L_A(a,c)$, thus these union $\mathsf{O}_3 = L_A(a,b) \cup L_A(a,c)$ is also REG-measurable by Lemma 19. The case O_4 can be archived in the same manner. ◀

Next we consider the so-called *Goldstine language*

$$\mathsf{G} \stackrel{\text{def}}{=} \{a^{n_1} b a^{n_2} b \cdots a^{n_p} b \mid p \geq 1, n_i \neq i \text{ for some } i\}.$$

While G can be accepted by a non-deterministic pushdown automaton, its generating function is not algebraic [13] and thus it is an inherently ambiguous context-free language due to the well-known Chomsky–Schützenberger theorem stating that the generating function of every unambiguous context-free language is algebraic [6].

▶ **Theorem 27.** G *is* REG*-measurable.*

Proof. Let $A = \{a,b\}$. Observe that $\mathsf{G} \subseteq A^*b$ and $\overline{\mu}_{\mathrm{REG}}(\mathsf{G}) \leq \delta_A(A^*b) = 1/2$. Let

$$L_{\mathsf{G}} = \{u \in A^* \mid uA^*\{b\} \cap \overline{\mathsf{G}} = \emptyset\}$$

be the set of all forbidden prefixes of the complement $\overline{\mathsf{G}}$. For each $k \geq 1$, we define

$$L_k \stackrel{\text{def}}{=} \{uA^*\{b\} \mid u \in L_{\mathsf{G}} \cap A^k\}.$$

If a word u is in L_{G}, then by definition of L_{G}, uvb is always in G for any word v, thus $L_k \subseteq G$ holds for each k. Any word in $\overline{L_{\mathsf{G}}} = A^* \setminus L_{\mathsf{G}}$ is a prefix of the infinite word

$a^{n_1}ba^{n_2}ba^{n_3}b\cdots$ ($n_i = i$ for each $i \in \mathbb{N}$) thus $\#(L_{\mathsf{G}} \cap A^n) = \#(A^n) - 1$ holds for each $n \geq 1$. Hence we have

$$\delta_A(L_k) = \lim_{n\to\infty} \frac{\#(L_k \cap A^n)}{\#(A^n)} = \lim_{n\to\infty} \frac{(\#(A^k) - 1)\cdot \#(A^{n-k-1})}{\#(A^n)}$$
$$= (\#(A)^k - 1)\cdot \#(A)^{-k-1} = 2^{-1} - 2^{-k-1}.$$

This implies that $\delta_A(L_k)$ tends to $1/2$. Thus $\mu_{\mathrm{REG}}(\mathsf{G}) = 1/2$. ◀

In general, for an infinite word $w \in A^\omega$, the set

$$\mathrm{Copref}(w) \stackrel{\mathsf{def}}{=} A^* \setminus \{u \in A^* \mid u \text{ is a prefix of } w\}$$

is called the *coprefix language of* w. The proof of Theorem 27 uses a key property that G can be characterised by using the coprefix language of the infinite word $w = a^{n_1}ba^{n_2}ba^{n_3}b\cdots$ as $\mathsf{G} = \mathrm{Copref}(w) \cap \{a,b\}^*\{b\}$ which was pointed out in [1]. Thus by the same argument, we can say that any coprefix language L is REG-measurable ($\mu_{\mathrm{REG}}(L) = 1$, in particular).

For coprefix languages, the following nice "gap theorem" holds.

▶ **Theorem 28** (Autebert–Flajolet–Gabarro [1]). *Let $w \in A^\omega$ be an infinite word generated by an iterated morphism, i.e., $w = h(w) = h^\omega(a)$ for some monoid morphism $h : A^* \to A^*$ and letter $a \in A$. Then for the coprefix language $L = \mathrm{Copref}(w)$ there are only two possibilities:*
1. *L is a regular language.*
2. *L is an inherently ambiguous context-free language.*

This means that we can construct, by finding some suitable morphism h, many examples of inherently ambiguous context-free languages.

4.3 K: A Language with Transcendental Density

We now show the fact that the language K defined by Kemp [16] (recall that the definition of K appeared in Therem 15) is REG-measurable. We will actually show a more general result regarding the following type of languages.

▶ **Definition 29.** Let $L \subseteq A^*$ be a language and $c \notin A$ be a letter. We call the language $L\{c\}(A \cup \{c\})^*$ over $A \cup \{c\}$ *suffix extension of L by c.*

▶ **Theorem 30.** *The suffix extension $L' \subseteq (A \cup \{c\})^*$ of any language $L \subseteq A^*$ by $c \notin A$ is* REG*-measurable.*

Proof. Let $B = A \cup \{c\}$ and $k = \#(B)$. We first show that L' has a natural density. For any words $u, v \in L$ with $u \neq v$, two languages $u\{c\}B^*$ and $v\{c\}B^*$ are disjoint, and clearly

$$\#(u\{c\}B^* \cap B^n)/\#(B^n) = \#\left(u\{c\}B^{n-|u|-1}\right)/\#(B^n) = k^{n-|u|-1}/k^n = k^{-(|u|+1)}$$

holds for $n > |u|$ thus $\delta_B(u\{c\}B^*) = k^{-(|u|+1)}$. The natural density of L' is

$$\delta_B(L') = \lim_{n\to\infty} \frac{\#(L' \cap B^n)}{\#(B^n)} = \lim_{n\to\infty} \frac{\#\left(\bigcup_{w\in L}(w\{c\}B^* \cap B^n)\right)}{\#(B^n)}$$
$$= \lim_{n\to\infty} \frac{\sum_{w\in L} \#(w\{c\}B^* \cap B^n)}{\#(B^n)} = \lim_{n\to\infty} \sum_{w\in(L\cap A^{<n})} k^{-(|w|+1)}. \tag{7}$$

Because the sequence $(\sum_{w\in(L\cap A^{<n})} k^{-(|w|+1)})_{n\in\mathbb{N}}$ is non-decreasing and bounded above by 1, the limit (7) exists, say $\delta_B(L') = \alpha$.

For each $n \in \mathbb{N}$, the language $L_n \stackrel{\text{def}}{=} \bigcup_{w \in L \cap A^{<n}} w\{c\}B^*$ is regular (since $L \cap A^{<n}$ is finite), $L_n \subseteq L'$ and $\delta_B(L_n) = \sum_{w \in (L \cap A^{<n})} k^{-(|w|+1)}$. Hence $\underline{\mu}_{\mathrm{REG}}(L') = \alpha$. By similar argument, for each $n \in \mathbb{N}$, we can claim that the language $K_n \stackrel{\text{def}}{=} A^* \setminus \bigcup_{w \in \overline{L} \cap A^{<n}} w\{c\}B^*$ satisfies $K_n \supseteq L'$ and $\delta_B(K_n)$ tends to α if n tends to infinity. Thus $\mu_{\mathrm{REG}}(L') = \alpha$. ◀

Since K is the suffix extensions of the union $S_1 \cup S_2$ in Theorem 15, we have:

▶ **Corollary 31.** K *is* REG*-measurable.*

▶ Remark 32. Theorem 30 indicates that REG-measurability is a quite relaxed property in some sense: even for a non-recursively-enumerable language, its suffix extension is still non-recursively-enumerable but REG-measurable.

The same proof method works for the *prefix extension*, and the *infix extension* is also REG-measurable.

▶ **Theorem 33.** *Let $c \notin A$ and $A' = A \cup \{c\}$. The prefix extension $L' = A'^*\{c\}L$ of any language $L \subseteq A^*$ is* REG*-measurable. Also, the infix extension $L'' = A'^*\{c\}L\{c\}A'^*$ of any language $L \subseteq A^*$ is* REG*-measurable, $\mu_{\mathrm{REG}}(L'') = 0$ if $L = \emptyset$, $\mu_{\mathrm{REG}}(L'') = 1$ otherwise, in particular.*

Proof. The prefix extension of L is just the reverse of the suffix extension of L, the same proof method trivially works. For the infix extension $L'' = A'^*\{c\}L\{c\}A'^*$, if $L = \emptyset$ then L'' is also empty and thus $\mu_{\mathrm{REG}}(L'') = 0$. Further, if $L \neq \emptyset$ then there is a word $w \in L$ and thus $A'^*cwcA'^* \subseteq L''$ holds, which means that $\delta_{A'}(A'^*cwcA'^*) = 1$ by the infinite monkey theorem and we have $\mu_{\mathrm{REG}}(L'') = 1$. ◀

4.4 Languages with Full REG-Gap

In Section 4.1, we showed that the language $L_{\{a,b\}}(a,b)$ is REG-measurable. On the other hand, by the result of Eisman–Ravikumar [10], we will know that the closely related language

$$\mathsf{M} \stackrel{\text{def}}{=} \{w \in \{a,b\}^* \mid |w|_a > |w|_b\},$$

sometimes called the *majority language*, is not REG-measurable. This contrast is interesting.

▶ **Theorem 34** (Eisman–Ravikumar [10, 11])**.** *Let $A = \{a,b\}$ and $L \subseteq A^*$ be a regular language. Then $\mathsf{M} \subseteq L$ implies*

$$\limsup_{n\to\infty} \{\#\left(\overline{L} \cap A^n\right) / \#(A^n)\} = 0.$$

One can easily observe that $\limsup_{n\to\infty} \{\#\left(\overline{L} \cap \#(A^n)\right) / \#(A^n)\} = 0$ if and only if $\delta_A(\overline{L}) = 0$, which means that any regular superset of M is co-null. Thus the above theorem implies that both M and $\overline{\mathsf{M}}$ are REG$^+$-immune, hence we have:

▶ **Corollary 35.** M *has full* REG*-gap.*

By using the infinite monkey theorem and some probabilistic arguments, we can generalise the previous theorem as follows.

▶ **Theorem 36.** *For any $m \geq 1$, the following language over $A = \{a,b\}$*

$$\mathsf{M}_m \stackrel{\text{def}}{=} \{w \in A^* \mid |w|_a > m \cdot |w|_b\}$$

has full REG*-gap, and $\delta_A(\mathsf{M}_m) = 1/2$ if $m = 1$ otherwise $\delta_A(\mathsf{M}_m) = 0$.*

Proof. First we prove that any non-null regular language L can not be a subset of M_m. Let $\eta : A^* \to M$ be the syntactic morphism η and monoid M of L, and let $c = \max_{m \in M} \min_{w \in \eta^{-1}(m)} |w|$ (this is well-defined natural number since M is finite). By the infinite monkey theorem, L is not null implies that L has no forbidden word, and thus for the word b^{2c} there exist two words x and y such that $xb^{2c}y$ is in L. We can assume that $|x|, |y| \leq c$ without loss of generality by the definition of c, which implies $|xb^{2c}y|_a \leq |x| + |y| = 2c \leq |xb^{2c}y|_b$ hence $xb^{2c}y \notin \mathsf{M}_m$. Thus $L \not\subseteq \mathsf{M}_m$ and $\underline{\mu}_{\mathrm{REG}}(\mathsf{M}_m) = 0$. By using same argument, we can prove that $\overline{\mu}_{\mathrm{REG}}(\mathsf{M}_m) = 0$ and hence M_m has full REG-gap.

In the case $m = 1$, $\delta_A(\mathsf{M}_1) = \delta_A(\mathsf{M}) = 1/2$ is obvious. It is enough to show that $\delta_A(\mathsf{M}_2) = 0$ holds (since $\mathsf{M}_m \subseteq \mathsf{M}_2$ for any $m \leq 2$). Indeed, we have

$$\delta_A(\mathsf{M}_2) = \lim_{n \to \infty} \frac{\#(\{w \in A^n \mid |w|_a > 2|w|_b\})}{2^n} = \lim_{n \to \infty} \frac{\#(\{w \in A^n \mid |w|_a > 2n/3\})}{2^n}$$
$$= \lim_{n \to \infty} \Pr(|\overline{X}_n - n/2| > n/6) = 0$$

where $\Pr(|\overline{X}_n - n/2| > n/6)$ means the probability that the absolute value of the difference of the number $\overline{X}_n$ of the occurrences of a's in a randomly chosen word of length n and its mean value $n/2$ is larger than $n/6$; its tends to zero by the weak law of large numbers. ◀

5 REG-Immesurability of Primitive Words

A non-empty word $w \in A^+$ is said to be primitive if $u^n = w$ implies $u = w$ for any $u \in A^+$ and $n \in \mathbb{N}$. The set of all primitive words over A is denoted by Q_A. Because the case $\#(A) = 1$ is meaningless ($\mathsf{Q}_A = A$ in this case), hereafter we always assume $\#(A) \geq 2$. Whether Q_A is context-free or not is a well-known long-standing open problem posed by Dömösi, Horváth and Ito [9]. Reis and Shyr [19] proved $\mathsf{Q}_A^2 = A^+ \setminus \{a^n \mid a \in A, n \neq 2\}$, which intuitively means that every non-empty word w not a power of a letter is a product of two primitive words. From this result one may think that Q_A is "very large" in some sense. Actually, Q_A is somewhat "large" (it is dense in the sense of Definition 9), but we can show more stronger property as follows.

▶ **Theorem 37.** $\delta_A(\mathsf{Q}_A) = 1$.

Proof. It is enough to show that $\delta_A(\overline{\mathsf{Q}_A}) = 0$ holds. One can easily observe that any natural number $n \in \mathbb{N}$ has at most $2\sqrt{n}$ divisors. In addition, for any non-primitive word $w = v^m$ of length n is uniquely determined by v (since $m = n/|v|$) and $|v| \leq n/2$. Hence the number of non-primitive words of length n satisfies

$$\#\left(\overline{\mathsf{Q}_A} \cap A^n\right) \leq 2\sqrt{n} \sum_{i=0}^{\lfloor n/2 \rfloor} \#\left(A^i\right) \leq 2\sqrt{n} \cdot \#(A)^{\lfloor n/2 \rfloor + 1}.$$

By using the above estimation, we can deduce that

$$\frac{\#\left(\overline{\mathsf{Q}_A} \cap A^n\right)}{\#(A^n)} \leq \frac{2\sqrt{n} \cdot \#(A)^{\lfloor n/2 \rfloor + 1}}{\#(A)^n} \leq \frac{2\sqrt{n}}{\#(A)^{n/2-1}}$$

and it tends to 0 if n tends to infinity (since we assume $\#(A) \geq 2$). Thus $\delta_A(\overline{\mathsf{Q}_A}) = 0$. ◀

While Q_A is "very large" (co-null) as stated above, we can also prove that Q_A is REG$^+$-immune. The proof relies on an analysis of the structure of the syntactic monoid of a non-null regular language. We assume that the reader has a basic knowledge of semigroup theory (*cf.* [18]): Green's relations $\mathcal{J}, \mathcal{R}, \mathcal{L}, \mathcal{H}$ and a direct consequence of Green's theorem (an $\mathcal{H}$-class H in a semigroup S is a subgroup of S if and only if H contains an idempotent), in particular.

▶ **Theorem 38.** *Any non-null regular language contains infinitely many non-primitive words, and hence* $\underline{\mu}_{\mathrm{REG}}(\mathsf{Q}_A) = 0$.

Proof. Let L be a regular language over A with a positive density $\delta_A(L) > 0$. We consider $\eta : A^* \to M$ the syntactic morphism η and the syntactic monoid M of L, and let S be a subset of M satisfying $\eta^{-1}(S) = L$. L is regular means that M is finite, and hence M has at least one $\leq_{\mathcal{J}}$-minimal element.

We first show that S contains a $\leq_{\mathcal{J}}$-minimal element t. This is rather clear because, for any non-$\leq_{\mathcal{J}}$-minimal element s, its language $\eta^{-1}(s) \subseteq A^*$ is null: s is non-$\leq_{\mathcal{J}}$-minimal means that there is an other element t such that $t <_{\mathcal{J}} s$ (*i.e.*, $MtM \subsetneq MsM$), whence $s \notin MtM$ which implies that any word $w \in \eta^{-1}(t)$ is a forbidden word of $\eta^{-1}(s)$. Thus by the infinite monkey theorem $\eta^{-1}(s)$ is null.

Clearly, we have $t^n \leq_{\mathcal{J}} t$ and thus $t \,\mathcal{J}\, t^n$ holds for any $n > 1$ by the $\leq_{\mathcal{J}}$-minimality of t. $t \,\mathcal{J}\, t^n$ implies that there is a pair of words x, y such that $xt^n y = t$. Since M is finite, x^m is an idempotent for some $m > 0$ (*i.e.*, $x^{2m} = x^m$). Thus we obtain $t = xt^n y = x(t)t^{n-1}y = x^2(t)(t^{n-1}y)^2 = \cdots = x^m t(t^{n-1}y)^m = x^m x^m t(t^{n-1}y)^m = x^m t$ whence $t = t^n(y(t^{n-1}y)^{m-1})$. It follows that $t \,\mathcal{R}\, t^n$. Dually, we also obtain $t \,\mathcal{L}\, t^n$ and hence we can deduce that $t \,\mathcal{H}\, t^n$ holds. By the finiteness of M, there exists some $n > 0$ such that t^n is an idempotent. Thanks to Green's theorem, the $\mathcal{H}$-equivalent class H_t of t is a subgroup of M with the identity element t^n. Because η is surjective, we can take a word w' from $\eta^{-1}(t)$. Let $t' = \eta(w'a) = t\eta(a)$ for some letter $a \in A$, then by the $\leq_{\mathcal{J}}$-minimality of t, we can take some words $x, y \in A^*$ so that $\eta(xw'ay) = \eta(x)t'\eta(y) = t$. Hence we can deduce that $\eta^{-1}(t)$ contains a non-empty word $w = xw'ay$. Then for any $\varepsilon \neq w \in \eta^{-1}(t)$ and $m \geq 1$, we have

$$\eta(w^{mn+1}) = t^{mn+1} = (t^n)^m \cdot t = t \in S$$

which means that $L \supseteq \eta^{-1}(t)$ contains infinitely many non-primitive words w^{mn+1}. ◀

▶ **Corollary 39** (of Theorem 37 and 38)**.** Q_A *has full* REG*-gap.*

▶ Remark 40. We emphasise that the assumption "L is non-null" in Theorem 38 is quite tight, since a slightly weaker assumption "L is of exponential growth" (*i.e.*, $\#(L \cap A^n)$ is exponential for n) does not imply that L contains non-primitive words. A trivial counterexample is $L_0 = \{a, b\}^* c$ over $A = \{a, b, c\}$: $\#(L_0 \cap A^n) = 2^{n-1}$ $(n \geq 1)$ is exponential but L_0 only consists of primitive words. L_0 has a cc as a forbidden word, hence it is null by the infinite monkey theorem. Thus L_0 is not a counterexample of Theorem 38.

6 Conclusion and Open Problems

In this paper we proposed REG-measurability and showed that several context-free languages are REG-measurable, excluding M_m. Interestingly, it is shown that, like G and K, languages that have been considered as complex from a combinatorial viewpoint are, actually, easy to asymptotically approximate by regular languages. It is also interesting that a modified majority language M_2 is just a deterministic context-free but it is complex from a measure theoretic viewpoint. Its complement $\overline{\mathsf{M}_2}$ is also deterministic context-free, and actually it is co-null but REG$^+$-immune (*i.e.*, has full REG-gap). This means that $\overline{\mathsf{M}_2}$ is as complex as Q_A from a viewpoint of REG-measurability.

The following fundamental problems are still open and we consider these to be future work.

▶ **Problem 41.** *Can we give an alternative characterisation of the null (resp. co-null) context-free languages (like Theorem 10)?*

▶ **Problem 42.** *Can we give an alternative characterisation of the* REG*-measurable context-free languages?*

▶ **Problem 43.** *Can we find a language class that can "separate"* Q_A *and* CFL*? i.e., Is there* $\mathcal{C}$ *such that* Q_A *has full* $\mathcal{C}$*-gap but no co-null context-free language has full* $\mathcal{C}$*-gap, or* Q_A *is* $\mathcal{C}$*-immeasurable but any co-null context-free language is* $\mathcal{C}$*-measurable?*

The our results (Theorem 36, 37 and 38) tell us that the class REG of regular languages can not separate Q_A and CFL. However, it is still open whether the situation is same or not when $\mathcal{C} = \mathrm{DetCFL}, \mathrm{UnCFL}$ or other extension of regular languages. Notice that *if* the answer of Problem 43 is "yes", then Q_A is not context-free.

Acknowledgement: The author would like to thank Takanori Maehara (RIKEN AIP) whose helpful discussion were an enormous help to me. The author also thank to anonymous reviewers for many valuable comments. This work was supported by JSPS KAKENHI Grant Number JP19K14582.

References

1 Jean-Michel Autebert, Philippe Flajolet, and Joaquim Gabarro. Prefixes of infinite words and ambiguous context-free languages. *Information Processing Letters*, 25(4):211 – 216, 1987.

2 Jean Berstel. Sur la densité asymptotique de langages formels. In *International Colloquium on Automata, Languages and Programming (ICALP, 1972)*, pages 345–358, France, 1973. North-Holland.

3 Jean Berstel, Dominique Perrin, and Christophe Reutenauer. *Codes and Automata*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2009.

4 Robert C. Buck. The measure theoretic approach to density. *American Journal of Mathematics*, 68(4):560–580, 1946.

5 Cezar Câmpeanu, Nicolae Sântean, and Sheng Yu. Minimal cover-automata for finite languages. *Theoretical Computer Science*, 267(1):3 – 16, 2001.

6 N. Chomsky and M.P. Schützenberger. The algebraic theory of context-free languages*. In *Computer Programming and Formal Systems*, volume 35, pages 118 – 161. Elsevier, 1963.

7 Brendan Cordy and Kai Salomaa. On the existence of regular approximations. *Theoretical Computer Science*, 387(2):125 – 135, 2007.

8 Michael Domaratzki. Minimal covers of formal languages. Master's thesis, University of Waterloo, 2001.

9 Pál Dömösi, Sándor Horváth, and Masami Ito. On the connection between formal languages and primitive words. pages 59–67, 1991.

10 Gerry Eisman and Bala Ravikumar. Approximate recognition of non-regular languages by finite automata. In *Twenty-Eighth Australasian Computer Science Conference (ACSC2005)*, volume 38 of *CRPIT*, Newcastle, Australia, 2005. ACS.

11 Gerry Eisman and Bala Ravikumar. On approximating non-regular languages by regular languages. *Fundamenta Informaticae*, 110:125–142, 2011.

12 Philippe Flajolet. Ambiguity and transcendence. In *Automata, Languages and Programming*, pages 179–188, Berlin, Heidelberg, 1985. Springer Berlin Heidelberg.

13 Philippe Flajolet. Analytic models and ambiguity of context-free languages. *Theoretical Computer Science*, 49(2):283 – 309, 1987.

14 Martin Kappes and Chandra M. R. Kintala. Tradeoffs between reliability and conciseness of deterministic finite automata. *Journal of Automata, Languages and Combinatorics*, 9(2–3):281–292, 2004.

15 Martin Kappes and Frank Nießner. Succinct representations of languages by dfa with different levels of reliability. *Theoretical Computer Science*, 330(2):299 – 310, 2005.

16 Rainer Kemp. A note on the density of inherently ambiguous context-free languages. *Acta Informatica*, 14(3):295–298, 1980.

17 Piet van Mieghem. *Graph Spectra for Complex Networks*. Cambridge University Press, 2010.

18 Jean-Éric Pin. Mathematical foundations of automata theory, 2012.

19 C.M. Reis and H.J. Shyr. Some properties of disjunctive languages on a free monoid. *Information and Control*, 37(3):334 – 344, 1978.

20 Arto Salomaa and Matti Soittola. *Automata Theoretic Aspects of Formal Power Series*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1978.

21 Ryoma Sin'ya. An automata theoretic approach to the zero-one law for regular languages: Algorithmic and logical aspects. In *Proceedings Sixth International Symposium on Games, Automata, Logics and Formal Verification, GandALF 2015*, pages 172–185, 2015.