A Note on Bucher's Density Problem*

Ryoma Sin'ya[a,*]

[a] *Akita University, 1-1 Tegata Gakuen-machi, Akita City, Akita Prefecture, 010-8502 JAPAN*

1. Introduction

For a set X, we denote by $\#(X)$ the cardinality of X. The complement of X is denoted by $\overline{X}$. We write $L_1 \subset_{\inf} L_2$ if $L_1 \subset L_2$ and $\#(L_2 \setminus L_1) = \infty$. We write DCFL, UCFL and CFL for the class of all deterministic context-free, unambiguous context-free and context-free languages.

Problem 1 (Bucher's density problem [1])**.** *Is the relation* $\subset_{\inf}$ *dense on* CFL*? Namely, for any* $L_1, L_2 \in$ CFL *with* $L_1 \subset_{\inf} L_2$*, is there always exists an intermediate* $L_3 \in$ CFL *such that* $L_1 \subset_{\inf} L_3 \subset_{\inf} L_2$*?*

This problem is still open. Actually, less attention has been paid to this problem posed by one-page-report [1] in 1980, perhaps due to the difficulty of the theory of context-free languages. The author can found very few published literature mentioning this problem [2, 3, 4].

One approach for this problem is finding a regular language R that *dissects* the margin $L_1 \setminus L_2$, *i.e.*, $\#((L_2 \setminus L_1) \cap R) = \infty$ and $\#\left((L_2 \setminus L_1) \cap \overline{R}\right) = \infty$. If there is such regular language R, we can set $L_3 = L_1 \cup (L_2 \cap R)$ so that $L_1 \subset_{\inf} L_3 \subset_{\inf} L_2$: this follows by the definition of R, and L_3 is context-free since the class of context-free languages is closed under intersection with regular languages and union. Yamakami and Kato [4] investigated the dissecting power of regular languages, and showed the following simple sufficient condition when a language is dissectable by regular languages. A language L is said to have the *constant growth property* if there exists $p, c > 0$ such that, for every $x \in L$ with $|x| \geq p$, there exists $y \in L$ such that $|x| < |y| < |x| + c$.

Lemma 1 ([4])**.** *Every infinite language having constant growth property is dissectable by some regular language.*

In this note, we show that if a language class $\mathcal{C}$ equips well-behaved generating functions and some closure properties, the relation $\subset_{\inf}$ is dense on $\mathcal{C}$.

*This document is the results of the research project funded by JSPS KAKENHI Grant Number JP19K14582.

*Corresponding author

Email address: ryoma@math.akita-u.ac.jp (Ryoma Sin'ya)

As a corollary, it will be shown that there always exists a context-free language L_3 in Problem 1 if L_1 and L_2 are both unambiguous. This is a stronger result than the previous best-known partial answer to this question stating that $\subset_{\inf}$ is dense on DCFL which is pointed out by Domaratzki [2] *without proof*. We also discuss about some hardness of this approach to general context-free languages and pose some problems.

2. Languages with Holonomic Generating Functions

In formal language theory, it is well-known that the generating function $F_L(x) = \sum_{n=0}^{\infty} \#\left(L \cap A^n\right) x^n$ of any unambiguous context-free language L is algebraic; this fact is called Chomsky-Schützenberger theorem [5]. Here a function F is algebraic (over $\mathbb{Q}$) means there exists a non-zero polynomial function $p(x, y)$ with rational coefficients that satisfies $p(x, F(x)) = 0$. We shall use more broader class of functions. A sequence $(a_n)_{n=0}^{\infty}$ is said to be *holonomic (of order r)* (over $\mathbb{Q}$) if there exist polynomials $p_0(x), \ldots, p_r(x)$ with rational coefficients with $p_0(x) \neq 0 \neq p_r(x)$ such that

$$p_0(n)a_n + p_1(n)a_{n+1} + \cdots + p_r(n)a_{n+r} = 0 \tag{1}$$

for all $n \in \mathbb{N}$ with $p_r(n) \neq 0$. A function F is said to be holonomic if it can be represented as $F = \sum_{n=0}^{\infty} a_n x^n$ for some holonomic sequence $(a_n)_{n=0}^{\infty}$. It is well-known that any algebraic function is holonomic (*cf.* [6]).

The following lemma is elemental and folklore, but we give a proof for self-containedness.

Lemma 2. *Let $(a_n)_{n=0}^{\infty}$ be a holonomic sequence with infinitely many non-zero points, i.e., $\#\left(\{n \mid a_n \neq 0\}\right) = \infty$. Then there exists $c > 0$ such that, for any $n \leq 0$ with $a_n \neq 0$, there exists $n < m < n + c$ such that $a_m \neq 0$.*

Proof. Suppose that there is no such constant c. Because $(a_n)_{n=0}^{\infty}$ is holonomic, there exists polynomials $p_0(x), \ldots, p_r(x)$ with $p_0(x) \neq 0 \neq p_r(x)$ such that (1) holds for all $n \in \mathbb{N}$ with $p_r(n) \neq 0$. By assumption, there is some $n > 0$ with $a_{n-1} \neq 0$ such that $a_n, a_{n+1}, \ldots, a_{n+r-1}$ are all zero. But this implies

$$p_0(n)a_n + p_1(n)a_{n+1} + \cdots + p_{r-1}(n)a_{n+r-1} + p_r(n)a_{n+r} = p_r(n)a_{n+r} = 0.$$

Thus $a_{n+r} = 0$ since $p_r(n) \neq 0$ by definition. Repeating this process, we obtain $a_m = 0$ for all $m \geq n$, which contradicts with $\#\left(\{n \mid a_n \neq 0\}\right) = \infty$. □

The above lemma says exactly that a language with holonomic generating function has the constant growth property. Further, if two functions F, G are holonomic, then $F - G$ is still holonomic (*cf.* [6]). This means that if we have two languages $L_1 \subset L_2$ whose generating functions are F and G, respectively, then the language $L_2 \setminus L_1$, whose generating function $G - F$ is holonomic, is dissectable by some regular language R by Lemma 2 and Lemma 1. Because $L_3 = L_1 \cup (L_2 \cap R)$ satisfies $L_1 \subset_{\inf} L_3 \subset_{\inf} L_2$, we have the following result.

Theorem 1. *Let $\mathcal{C}$ be a class of languages closed under intersection with regular languages and union. If any language in $\mathcal{C}$ has a holonomic generating function, then the relation $\subset_{\inf}$ is dense on $\mathcal{C}$.*

Unfortunately, UCFL is not closed under union ($\{a^i b^i c^j \mid i,j \in \mathbb{N}\} \cup \{a^i b^j c^j \mid i,j \in \mathbb{N}\} \notin$ UCFL for example), but we have the following as a corollary.

Corollary 1. *For any $L_1, L_2 \in$ UCFL with $L_1 \subset_{\inf} L_2$, there exists $L_3 \in$ CFL such that $L_1 \subset_{\inf} L_3 \subset_{\inf} L_2$.*

Another example of languages with holonomic generating functions are languages recognised by *weakly-unambiguous Parikh automata* [7]. We omit the detailed definition here, but only notice that the class WUPA of all such languages is incomparable with CFL under the inclusion relation.

3. Discussion and Open Problems

In general, a language without holonomic generating function does not satisfy the constant growth property. The following proposition gives a simple counterexample in CFL.

Proposition 1. *There exists a pair of context-free languages L_1, L_2 such that $L_1 \subset_{\inf} L_2$ does not have the constant growth property.*

Proof. Let $A = \{a, b\}$, and let G be the *Goldstein language* defined by

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

It is well-known that G is (inherently ambiguous) context-free [8]. Define $G' = G \cup A^* a$. One can easily observe that

$$\overline{G'} = A^* \setminus G' = \{a^1 b a^2 b b \cdots a^n b \mid n \geq 1\}$$

holds (hence $G' \subset_{\inf} A^*$) and $\overline{G'}$ does not have the constant growth property. □

In other words, $\overline{G'}$ is very "sparse". From this viewpoint, one may think that $\overline{G'}$ can not be dissectable by regular languages. However, actually, it can: observe that the length of $a^1 b a^2 b b \cdots a^n b \in \overline{G'}$ is $n + n(n+1)/2$, which is even when n is a multiple of 4 and odd when $n = 2m$ for odd m. Thus $\overline{G'}$ contains infinitely many words with both even and odd length, hence $(AA)^*$ dissects it.

The Goldstein language above is closely related to more general notion *co-prefix languages*. For an infinite word $\alpha \in A^\omega$, we define the co-prefix language of α as

$$\mathrm{Copref}(\alpha) = A^* \setminus \{w \in A^* \mid w \text{ is a prefix of } \alpha\}.$$

Then G can be represented as $\mathrm{Copref}(\alpha_1) \cap A^* b$ where

$$\alpha_1 = a^1 b a^2 b a^3 \cdots a^n b \cdots .$$

The above α_1 is some simple kind of infinite words. If $h : A^* \to A^*$ is a monoid morphism and $h(a) = aw$ for some $w \in A^*$, then there uniquely exists an infinite word α satisfies

$$h(\alpha) = \alpha = h^\omega(a).$$

Such word α is said to be *generated by an iterated morphism*. Consider the following morphism h_1 and g_1 over $A = \{a, b, c\}$:

$$h_1(a) = abc \qquad h_1(b) = bc \qquad h_1(c) = c$$
$$g_1(a) = \varepsilon \qquad g_1(b) = a \qquad g_1(c) = b$$

Then we have

$$h_1(a) = abc, h_1^2(a) = h_1(abc) = abcbcc,$$
$$h_1^3(a) = h_1(abcbcc) = abcbccbccc, \cdots,\ h_1^n(a) = abcbccbccc \cdots bc^n$$

and hence $g_1(h_1^\omega(a)) = \alpha_1$. Berstel [9] showed that a co-prefix language of the infinite word generated by an iterated morphism is always context-free, and Flajolet et al. [10] further showed that such language is either regular or inherently ambiguous.

By using the co-prefix construction, Honkala [11] construct a context-free language L whose complement $\overline{L}$ is more sparser than $\overline{G'}$, *i.e.*, the lengths of words in L is of exponential growth $2^{n-1} + n$ (see Corollary 5.7.3 of [12] for the detailed construction). However, $2^{n-1} + n$ is odd if $n \geq 3$ is odd and even otherwise, thus $(AA)^*$ still dissects $\overline{L}$.

By slightly modifying the construction of L, we can obtain a context-free language K such that $\overline{K}$ is of exponential growth and the length of every word in $\overline{K}$ is even as follows. Let $A = \{a, b, c\}$ and

$$h_2(a) = abc \qquad h_2(b) = b^2 \qquad h_2(c) = c^2$$

be a morphisms on A. Observe that

$$h_1(a) = abc, h_1^2(a) = h_1(abc) = abcb^2c^2,$$
$$h_1^3(a) = h_1(abcb^2c^2) = abcb^2c^2b^4c^4, \cdots,$$
$$h_2^n(a) = abcb^2c^2b^4c^4 \cdots b^{2^{n-1}}c^{2^{n-1}}$$

holds. Let $\alpha = h(\alpha)$ be the infinite word generated by h. Define

$$K = \mathrm{Copref}(\alpha) \cup A^*\{bb, cb, cc\} \cup \{\varepsilon, a, ab\}.$$

By construction, $\overline{K} = \{abcb^2c^2b^4c^4 \cdots b^nc \mid n \geq 1\}$ thus

$$\{|w| \mid w \in \overline{K}\} = \{1 + (2^{n+1} - 1) + (2^n - 1) + 1 \mid n \geq 1\} = \{3 \cdot 2^n \mid n \geq 1\}.$$

Thus $(AA)^*$ can not dissect $\overline{K}$. But clearly, a regular language like $a((b^*c^*)^2)^*$ can dissect it:

$$\overline{K} \cap a((b^*c^*)^2)^* = \{abcb^2c^2b^4c^4 \cdots b^nc \mid n \geq 1 \text{ and } n \text{ is even }\}.$$

It is still open that can we construct a context-free language L whose complement is not dissectable by any regular language. If it is possible, then the co-prefix construction might be useful for giving such examples.

References

[1] Bucher, W.: A density problem for context-free languages. Bull. Eur. Assoc. Theor. Comput. Sci (10) (1980) 53

[2] Domaratzki, M.: Minimal covers of formal languages. Master's thesis, University of Waterloo (2001)

[3] Shallit, J.: A Second Course in Formal Languages and Automata Theory. 1 edn. Cambridge University Press, USA (2008)

[4] T. Yamakami, Y.K.: The dissecting power of regular languages. Inf. Process. Lett (113) (2013) 116–122

[5] Chomsky, N., Schützenberger, M.: The algebraic theory of context-free languages*. In: Computer Programming and Formal Systems. Volume 35. Elsevier (1963) 118–161

[6] Kauers, M., Paule, P.: The Concrete Tetrahedron - Symbolic Sums, Recurrence Equations, Generating Functions, Asymptotic Estimates. Texts & Monographs in Symbolic Computation. Springer (2011)

[7] Bostan, A., Carayol, A., Koechlin, F., Nicaud, C.: Weakly-unambiguous parikh automata and their link to holonomic series. In Czumaj, A., Dawar, A., Merelli, E., eds.: 47th International Colloquium on Automata, Languages, and Programming, ICALP 2020, July 8-11, 2020, Saarbrücken, Germany (Virtual Conference). Volume 168 of LIPIcs., Schloss Dagstuhl - Leibniz-Zentrum für Informatik (2020) 114:1–114:16

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

[9] Berstel, J.: Every iterated morphism yields a co-cfl. Information Processing Letters **22**(1) (1986) 7–9

[10] Autebert, J.M., Flajolet, P., Gabarro, J.: Prefixes of infinite words and ambiguous context-free languages. Information Processing Letters **25**(4) (1987) 211–216

[11] Honkala, J.: On a problem of g. păun. Bull. Eur. Assoc. Theor. Comput. Sci (64) (1998) 341

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