Measuring Power of Generalised Definite Languages

Ryoma Sin'ya

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

Abstract. A language L is said to be $\mathcal{C}$-measurable, where $\mathcal{C}$ is a class of languages, if there is an infinite sequence of languages in $\mathcal{C}$ that "converges" to L. In this paper, we investigate the measuring power of GD of the class of all generalised definite languages. Although each generalised definite language only can check some local property (prefix and suffix of some bounded length), it is shown that many non-generalised-definite languages are GD-measurable. Further, we show that it is decidable whether a given regular language is GD-measurable or not.

1 Introduction

$\mathcal{C}$-measurability for a class $\mathcal{C}$ of languages is introduced by [14] and it was used for classifying non-regular languages by using regular languages. A language L is said to be $\mathcal{C}$-measurable if there is an infinite sequence of languages in $\mathcal{C}$ that converges to L. Roughly speaking, L is $\mathcal{C}$-measurable means that it can be approximated by a language in $\mathcal{C}$ with *arbitrary high precision*: the notion of "precision" is formally defined by the density of formal languages. Hence that a language L is not $\mathcal{C}$-measurable ($\mathcal{C}$-*im*measurable) means that L has a complex shape so that it can not be approximated by languages in $\mathcal{C}$. While the membership problem for a given language L and a class $\mathcal{C}$ just asks whether $L \in \mathcal{C}$, the $\mathcal{C}$-measurability asks the existence of an infinite sequence of languages in $\mathcal{C}$ that converges to L. In this sense, measurability is much more difficult than the membership problem and its analysis is a challenging task. For example, the author [15] showed that, for the class SF of all star-free languages, the class of all SF-measurable regular languages strictly contains SF but does not contain some regular languages. However, the decidability of SF-measurability for regular languages is still unknown. Only for some very restricted subclasses $\mathcal{C}$ of star-free languages, the decidability of $\mathcal{C}$-measurability is known [16]. A language L is called *locally testable* [5,9,18] if it is a finite Boolean combination of languages of the form uA^*, A^*v and A^*wA^*. Although the definition of local testability is very simple, it was shown in [16] that many non-locally-testable languages are LT-measurable, where LT is the class of all locally testable languages, and any *unambiguous polynomial* (language definable by the first-order logic with two variables) is LT-measurable. However, the decidability of LT-measurability for regular languages was left open in [16].

In this paper, as a continuation research of [16], we examine the measuring power of languages defined by *definiteness*, which is a natural restriction of the notion of local testability. A language L is called *definite* (*reverse definite*, respectively) [3] if it is a finite Boolean combination of languages of the form A^*u (uA^*, respectively). Also, L is called *generalised definite* [7] if it is a finite Boolean combination of languages of the form uA^* and A^*v. We consider GD-measurability and also consider D-measurability and RD-measurability where D, RD and GD is the class of all definite, reverse definite and generalised definite languages. The main results of this paper are two folds. We show:

(1) A simple automata theoretic and algebraic characterisation of RD-measurability (Theorem 1 and Theorem 3).
(2) The equivalence of the GD-measurability and the LT-measurability (Proposition 1) and a decidable characterisation of GD-measurability (Theorem 4). This decidability result answers a question posed in [16].

The structure of this paper is as follows. Section 2 provides preliminaries including density, measurability and definitions of fragments of locally testable languages. An automata theoretic characterisation of RD-measurability is given in Section 3, and a decidable characterisation of GD-measurability is given in Section 4, respectively. Related and future work are described in Section 5.

2 Preliminaries

This section provides the precise definitions of density, measurability and local varieties of regular languages. REG_A denotes the family of all regular languages over an alphabet A. We assume that the reader has a standard knowledge of automata theory including the concept of syntactic monoids (*cf.* [8]).

2.1 Languages and automata

For an alphabet A, we denote the set of all words (all non-empty words, respectively) over A by A^* (A^+, respectively). We write $|w|$ for the length of w and A^n for the set of all words of length n. For a word $w \in A^*$ and a letter $a \in A$, $|w|_a$ denotes the number of occurrences of a in w. We denote by $w^r = a_k \cdots a_1$ the reverse of $w = a_1 \cdots a_k$, and denote by $L^r = \{w^r \mid w \in L\}$ the reverse of the language L. A word v is said to be a factor 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. 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 w is called a *forbidden word* of L.

A deterministic automaton $\mathcal{A}$ over A is a quadruple $\mathcal{A} = (Q, \cdot, q_0, F)$ where Q is a finite set of states, $\cdot : Q \times A \to Q$ is a transition function, $q_0 \in Q$ is an initial and $F \subseteq Q$ is a set of final states. The language recognised by $\mathcal{A}$ is denoted by $L(\mathcal{A}) = \{w \in A^* \mid q_0 \cdot w \in F\}$. For a set of states $Q' \subseteq Q$ and a word w, we write $Q' \cdot w$ for the set of transition states from Q' by w: $Q' \cdot w = \{q \cdot w \mid q \in Q'\}$.

The automaton $\mathcal{A}$ is called accessible if for every state $p \in Q$ there is a word w such that $q_0 \cdot w = p$. In this paper, we only consider accessible deterministic automata. Q' is called strongly connected if for every $p, q \in Q'$, there is some word w such that $p \cdot w = q$. We say that Q' is a *sink* if it is strongly connected and there is no outgoing transition from Q', *i.e.*, $Q' \cdot w \subseteq Q'$ for any w.

2.2 Locally testable and definite languages

For a family $\mathcal{C}_A$ of languages over A, we denote by $\mathscr{B}\mathcal{C}_A$ the finite Boolean closure of $\mathcal{C}_A$. The class LT_A of all locally testable languages over A can be defined as

$$\mathrm{LT}_A = \mathscr{B}\{wA^*, A^*w, A^*wA^* \mid w \in A^*\}.$$

The class $\mathrm{D}_A, \mathrm{RD}_A$ and GD_A of all *definite, reverse definite* [3] and *generalised definite* [7] languages over A are defined as follows:

$$\mathrm{D}_A = \mathscr{B}\{A^*w \mid w \in A^*\}, \qquad \mathrm{RD}_A = \mathscr{B}\{wA^* \mid w \in A^*\},$$
$$\mathrm{GD}_A = \mathscr{B}\{A^*w, wA^* \mid w \in A^*\}.$$

Hence these classes are proper subclasses of locally testable languages.

Remark 1 (cf. [5]). In [3,7] definite languages are originally defined as follows. A language L is called:

- definite if and only if $L = E \cup A^*F$ for some finite sets $E, F \subseteq A^*$.
- reverse definite if and only if $L = E \cup FA^*$ for some finite sets $E, F \subseteq A^*$.
- generalised definite if and only if $L = E \cup \bigcup_{i \in I} F_i A^* G_i$ for some finite sets E and $F_i, G_i \subseteq A^*$ for all $i \in I$, where I is a finite index set.

For any word $w \in A^*$, the singleton $\{w\}$ can be written as the Boolean combination $wA^* \cap \overline{\bigcup_{a \in A} waA^*}$, hence any finite subset $F \subseteq A^*$ is in $\mathscr{B}\{wA^* \mid w \in A^*\}$. Conversely, for any w, the complement $\overline{wA^*}$ can be written in the form of a reverse definite language: $\{u \in A^* \mid |u| < |w|\} \cup (A^{|w|} \setminus \{w\})A^*$. Hence, these original definitions can be modified by using the finite Boolean closure as above.

2.3 Density and measurability 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.

Definition 1 (*cf.* [2]). The *density* $\delta_A(L)$ of $L \subseteq A^*$ is defined as

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

if it exists, otherwise we write $\delta_A(L) = \perp$. The language L is called *null* if $\delta_A(L) = 0$, and dually, L is called *co-null* if $\delta_A(L) = 1$.

Example 1. It is known that every regular language has a rational density (*cf.* [11]) and it is computable. Here we explain two examples of (co-)null languages.

(1) For each word w, the language A^*wA^*, the set of all words that contain w as a factor, is of density one (co-null). This fact follows from the so-called the *infinite monkey theorem* (this is also called as "Borges's theorem", *cf.* [6, p.61, Note I.35]): take any word w. A random word of length n contains w as a factor with probability tending to 1 as $n \to \infty$.
A language L having a forbidden word w is always null: having a forbidden word w means $A^*wA^* \subseteq \overline{L}$ hence we have $\delta_A(A^*wA^*) \leq \delta_A(\overline{L})$, which implies $\delta_A(\overline{L}) = 1$ by the infinite monkey theorem.
(2) The set of all palindromes $L_{\text{pal}} = \{w \in A^* \mid w = w^r\}$ over $A = \{a, b\}$ is dense but null. This follows from the fact that $\#(L_{\text{pal}} \cap A^n)$ equals to $2^{\lceil n/2 \rceil}$ and $2^{\lceil n/2 \rceil}/2^n < 2^{(1-n/2)}$ tends to zero if n tends to infinity.

We list some basic properties of the density as follows.

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 $K \subseteq L$. (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 $K \cap L = \emptyset$. (6) $\delta_A(uL) = \delta_A(Lu) = \delta_A(L) \cdot \#(A)^{-|u|}$ for each $u \in A^*$.*

For more properties of δ_A, see Chapter 13 of [2].

The notion of "measurability" on formal languages is defined by a standard measure theoretic approach as follows.

Definition 2 ([14]). Let $\mathcal{C}_A$ be a family of languages over A. For a language $L \subseteq A^*$, we define its $\mathcal{C}_A$*-inner-density* $\underline{\mu}_{\mathcal{C}_A}(L)$ and $\mathcal{C}_A$*-outer-density* $\overline{\mu}_{\mathcal{C}_A}(L)$ over A as

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

A language L is said to be $\mathcal{C}_A$*-measurable* if $\underline{\mu}_{\mathcal{C}_A}(L) = \overline{\mu}_{\mathcal{C}_A}(L)$ holds. 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 $\lim_{n\to\infty} \delta_A(L_n) = \delta_A(L)$.

We give some examples of LT_A-(im)measurable languages from [14,16].

Example 2. (1) The set of all palindromes $L_{\text{pal}} = \{w \in A^* \mid w = w^r\}$ is LT_A-measurable. The sequence of locally testable languages $L_k = \{wA^*w^r \mid |w| = k\}$ converges to L_{pal} from outer if k tends to infinity (see [14] for the detail). The density of L_{pal} is zero as stated in Example 1, hence the constant sequence of the empty language trivially converges to L_{pal} from inner.
(2) For any real number $\alpha \in [0, 1]$, there is an LT_A-measurable language L whose density is α. See [15] for the detailed construction.
(3) The language $M_k = \{w \in \{a, b\}^* \mid |w|_a = |w|_b \mod k\}$ is LT-*im*measurable for any $k \geq 2$. See [16] or Section 4.1 for the proof.

For a family $\mathcal{C}_A$ of languages over A, we denote by $\mathrm{Ext}_A(\mathcal{C}_A)$ ($\mathrm{RExt}_A(\mathcal{C}_A)$, respectively) the class of all $\mathcal{C}_A$-measurable languages ($\mathcal{C}_A$-measurable *regular* languages, respectively) over A. A family of regular languages over A is called a *local variety* [1] over A if it is closed under Boolean operations and left-and-right quotients.

Lemma 2 ([15]). Ext_A *is a closure operator, i.e., it satisfies the following three properties for each* $\mathcal{C} \subseteq \mathcal{D} \subseteq 2^{A^*}$*: (extensive)* $\mathcal{C} \subseteq \mathrm{Ext}_A(\mathcal{C})$*, (monotone)* $\mathrm{Ext}_A(\mathcal{C}) \subseteq \mathrm{Ext}_A(\mathcal{D})$*, and (idempotent)* $\mathrm{Ext}_A(\mathrm{Ext}_A(\mathcal{C})) = \mathrm{Ext}_A(\mathcal{C})$*. Moreover,* RExt_A *is a closure operator over the class of all local varieties of regular languages over A, i.e.,* $\mathcal{C}_A$*-measurability is preserved under Boolean operations and quotients for any local variety* $\mathcal{C}_A$*.*

The following lemma is useful and will be used in Section 3 and Section 4.

Lemma 3. *Let* $\mathcal{A} = (Q, \cdot, q_0, F)$ *be a deterministic automaton,* $Q_1, \cdots, Q_k$ *be its all sink components and let* $Q' = Q \setminus \bigcup_{i=1}^{k} Q_i$*. Then the language* $P' = \{w \in A^* \mid q_0 \cdot w \in Q'\}$ *is of density zero,* $P_i = \{w \in A^* \mid q_0 \cdot w \in Q_i\}$ *satisfies* $P_i = P_i A^*$ *and has a non-zero density for each* i*.*

Proof. The condition $P_i = P_i A^*$ is clear because Q_i is a sink for each i: $Q_i \cdot w \subseteq Q_i$ holds for every w. For each i, P_i is non-empty because $\mathcal{A}$ is accessible (all automata in this paper are accessible as stated in Section 2.1). Let w be a word in P_i. By Lemma 1, we have $\delta_A(P_i) \geq \delta_A(wA^*) = \#(A)^{-|w|} > 0$, *i.e.*, P_i has a non-zero density. Now we show that the density of P' is zero. Let $Q' = \{q_0, q_1, \cdots, q_n\}$. For every state q_i in Q', there exists some word w_{q_i} such that $q_i \cdot w_{q_i}$ is in some sink component. Because every q_i in Q' is not in any sink component, q_i is not reachable from the state $q_i \cdot w_{q_i}$, *i.e.*$(q_i \cdot w_{q_i}) \cdot w \notin Q'$ for every w. Define $u_0 = w_{q_0}$ and $u_i = w_{q_i \cdot v_{i-1}}$ if $q_i \cdot v_{i-1} \in Q'$ and $u_i = \varepsilon$ otherwise for each $i \in \{1, \cdots, n\}$ where v_{i-1} is the word of the form $u_0 \cdots u_{i-1}$. By the construction, for every q_i in Q', we have $q_i \cdot u_0 \cdots u_n \notin Q'$. This means that $u_0 \cdots u_n$ is a forbidden word of P' and hence P' is of density zero by the infinite monkey theorem. $\square$

For simplicity, here after we fix an alphabet A and omit the subscript A for denoting local varieties.

3 Simple Characterisation of RD-Measurability

The next theorem gives a simple automata theoretic characterisation of RD-measurability.

Theorem 1. *For a minimal deterministic automaton* $\mathcal{A}$*, the followings are equivalent:*

(1) *Every sink component of* $\mathcal{A}$ *is a singleton.*
(2) $L(\mathcal{A})$ *is* RD*-measurable.*

Proof. Let $L = L(\mathcal{A}), Q_1, \cdots, Q_k$ be all sink components of $\mathcal{A} = (Q, \cdot, q_0, F)$ and let $Q' = Q \setminus \bigcup_{i=1}^{k} Q_i$. For each $i \in \{1, \cdots, k\}$, define $P_i = \{w \in A^* \mid q_0 \cdot w \in Q_i\}$ and define $P' = \{w \in A^* \mid q_0 \cdot w \in Q'\}$. Clearly, $P_1, \cdots, P_k$ and P' form the partition of A^*, and we have $\delta_A(P') = 0$ by Lemma 3.

Proof of (1) $\Rightarrow$ (2)**:** Because each Q_i is a singleton, P_i is contained in L if the state in Q_i belongs to F and P_i is contained in $\overline{L}$ otherwise. Define

$$M = \bigcup \{P_i \mid Q_i \subseteq F\} \quad \text{and} \quad M' = \bigcup \{P_i \mid Q_i \subseteq Q \setminus F\}.$$

By the definition and Lemma 3 , we have $M = MA^* \subseteq L$ and $M' = M'A^* \subseteq \overline{L}$. Because $P_1, \cdots, P_k, P'$ form the partition of A^* and the density of P' is zero, we can deduce that $\delta_A(M) + \delta_A(M') = 1$, which implies $\delta_A(M) = \delta_A(L)$ and $\delta_A(M') = \delta_A(\overline{L'})$. For each $n \in \mathbb{N}$ and i, the set $M_n = \{w \in M \mid |w| \leq n\}$ and $M'_n = \{w \in M' \mid |w| \leq n\}$ are finite and hence the sequence of reverse definite languages $M_n A^*$ and $\overline{M'_n A^*}$ converges to L from inner and outer, respectively, *i.e.*(i) $M_n A^* \subseteq L$ and $\overline{M'_n A^*} \supseteq L$ holds for each n, and (ii) $\lim_{n\to\infty} \delta_A(M_n A^*) = \delta_A(L)$ and $\lim_{n\to\infty} \delta_A(\overline{M'_n A^*}) = \delta_A(L)$.

Proof of (2) $\Rightarrow$ (1)**:** This direction is shown by contraposition. We assume that (1) is not true, *i.e.*, some sink component, say Q_j, is not a singleton. By the minimality of $\mathcal{A}$, Q_j contains at least one final state, say p, and at least one non-final state, say p' (if not, all states in Q_j are right equivalent). For each $q \in Q_j$, we write L_q for the language $L_q = \{w \in A^* \mid q_0 \cdot w = q\}$.

Because P_j is non-empty and $P_j = P_j A^*$ holds, the density of P_j is not zero. P_j has non-zero density implies that there exists at least one state q in P_j such that L_q has non-zero density. Since Q_j is a sink (strongly connected, especially), there exist some words $w_{q,p}$ and $w_{p,p'}$ such that $q \cdot w_{q,p} = p$ and $p \cdot w_{p,p'} = p'$. Thus $L_q w_{q,p} \subseteq L_p$ holds, from which we can deduce that $\delta_A(L_p) \geq \delta_A(L_q w_{q,p}) = \delta_A(L_q) \cdot \#(A)^{-|w_{q,p}|} > 0$, *i.e.*, L_p has non-zero density, say $\alpha > 0$.

We can show that, for every reverse definite language $R = E \cup FA^*$ (where E, F are finite sets) such that $R \subseteq L$, $FA^* \cap L_p = \emptyset$ holds as follows. If there is some word $w \in FA^* \cap L_p$, then $ww_{p,p'}$ is in $FA^* \cap \overline{L}$ since $ww_{p,p'} \in L_{p'}$ and p' is non-final. This violates the assumption $R \subseteq L$. This means that every reverse definite subset $R = E \cup FA^*$ of L should have density less than or equal to $\delta_A(L \setminus L_p) = \delta_A(L) - \alpha < \delta_A(L)$. Hence, no sequence of reverse definite languages converges to L from inner. $\square$

For a given automaton $\mathcal{A}$, we can construct its reverse automaton $\mathcal{A}^r$ recognising $L(\mathcal{A})^r$ by flipping final and non-final states and reversing transition relations. By the definition of definite and reverse definite languages, L is D-measurable if and only if L^r is RD-measurable. Hence, we can use Theorem 1 to deduce the decidability of D-measurability.

Corollary 1. *For a given regular language L it is decidable whether L is* RD-*measurable (*D*-measurable, respectively).*

3.1 Algebraic characterisation

In this subsection we give an algebraic characterisation of RD-measurability, which is a natural analogy of the algebraic characterisation of RD stated as follows. Let S be a semigroup. An element $x \in S$ is called a left zero if $xS = \{x\}$ holds. An element $x \in S$ is called an idempotent if $x^2 = x$ holds.

Theorem 2 (*cf.* [4])**.** *For a regular language L and its syntactic semigroup S_L, the followings are equivalent:*

(1) *L is in* RD*.*
(2) *Every idempotent of S_L is a left zero.*

Let M be a monoid. For elements x and y in M, we write $x \leq_{\mathcal{R}} y$ if $xM \subseteq yM$ holds. Notice that $x \leq_{\mathcal{R}} y$ if and only if $yz = x$ for some $z \in M$. An element x is called *$\mathcal{R}$-minimal* if $y \leq_{\mathcal{R}} x$ implies $x \leq_{\mathcal{R}} y$ for every y in M.

Theorem 3. *For a regular language L and its syntactic monoid M_L, the followings are equivalent:*

(1) *L is* RD*-measurable.*
(2) *Every $\mathcal{R}$-minimal element of M_L is a left zero.*
(3) *Every $\mathcal{R}$-minimal idempotent of M_L is a left zero.*

Proof. Let $\mathcal{A} = (Q, \cdot, q_0, F)$ be the minimal automaton of L. Notice that M_L is isomorphic to the transition monoid $T = (\{f_w : Q \to Q \mid w \in A^*\}, \circ, f_\varepsilon)$ of $\mathcal{A}$ where f_w is the map defined by $f_w(q) = q \cdot w$, the multiplication operation $\circ$ is the composition $f_u \circ f_v = f_{uv}$ and the identity element f_ε is the identity mapping on Q. Hence, we identify M_L with T.
Proof of (1) $\Rightarrow$ (2)**:** Let f be an $\mathcal{R}$-minimal element of T. If $f(q)$ is not in any sink component of $\mathcal{A}$ for some q, there is a some word w such that $(f \circ f_w)(q) = f(q) \cdot w$ is in some sink component. But this means that f is not $\mathcal{R}$-minimal because q is not reachable by $f(q)\cdot w$, which implies $(f \circ f_w) \circ g \neq f$ for any $g \in T$. Hence, $f(q)$ is in some sink component. By the assumption and Theorem 1, every sink component of $\mathcal{A}$ is a singleton. This means that $f(q) \cdot w = q$ holds for every w, *i.e.*, f is a left zero.
Proof of (2) $\Rightarrow$ (1)**:** This direction is shown by contraposition. Assume (1) is not true. That is, there is a sink component $Q' \subseteq Q$ which is not a singleton by Theorem 1. Let p and q in Q' be two different states and f be an $\mathcal{R}$-minimal element in T such that $f(q_0) = p$ (such f always exists since $\mathcal{A}$ is accessible and Q' is sink). Because Q' is strongly connected, there is some word w such that $p \cdot w = q$. This means that $f \neq f \circ f_w$ (because $f(q_0) = p \neq q = (f \circ f_w)(q_0)$), *i.e.*, f is not a left zero.
Proof of (2) $\Leftrightarrow$ (3)**:** (2) implies (3) is trivial. Assume (3). Let x be an $\mathcal{R}$-minimal element of M_L. Because M_L is finite, there is some index $i \geq 1$ such that x^i is an idempotent. By the $\mathcal{R}$-minimality of x and $x^i = x \cdot x^{i-1} \leq_{\mathcal{R}} x$, $x^i \cdot y = x$ holds for some y. But x^i is a left zero by the assumption, this means that $x = x^i$. $\square$

4 Decidable Characterisation of GD-Measurability

In this section we consider the GD-measurability. First we show that the GD-measurability is equivalent to the LT-measurability.

Proposition 1. *A language L is* LT*-measurable if and only if L is* GD*-measurable.*

Proof. For proving the equivalence $\mathrm{Ext}_A(\mathrm{LT}) = \mathrm{Ext}_A(\mathrm{GD})$, it is enough to show that every locally testable language is GD-measurable by the monotonicity and idempotency of Ext_A (Lemma 2): $\mathrm{Ext}_A(\mathrm{GD}) \supseteq \mathrm{LT}$ implies $\mathrm{Ext}_A(\mathrm{GD}) = \mathrm{Ext}_A(\mathrm{Ext}_A(\mathrm{GD})) \supseteq \mathrm{Ext}_A(\mathrm{LT}) \supseteq \mathrm{Ext}_A(\mathrm{GD})$. Further, since GD is closed under Boolean operations, GD-measurability is closed under Boolean operations by Lemma 2 and hence we only have to show that wA^*, A^*w and A^*wA^* are all GD-measurable for every w. The languages of the form wA^* and A^*w are already in GD, thus it is enough to show that A^*wA^* is GD-measurable. This was essentially shown in [16] as follows. Since the case $w = \varepsilon$ is trivial, we assume $w = a_1 \cdots a_n$ where $a_i \in A$ and $n \geq 1$. Define $W_k = (A^k \setminus K_k)wA^*$ where $K_k = \{u \in A^k \mid ua_1 \cdots a_{n-1} \in A^*wA^*\}$ for each $k \geq 0$. Intuitively, W_k is the set of all words in which w *firstly* appears at the position $k+1$ as a factor. By definition, W_k is generalised definite (reverse definite, in particular). Clearly, $W_i \cap W_j = \emptyset$ and $\delta_A(W_i) > 0$ for each $i \neq j$, thus we have $\bigcup_{k\geq 0} W_k = A^*wA^*$ and hence $\lim_{n\to\infty} \delta_A\left(\bigcup_{k\geq 0}^{n} W_k\right) = 1$, *i.e.*, $\underline{\mu}_{\mathrm{GD}}(A^*wA^*) = 1$. Thus $A^*wA^* \in \mathrm{Ext}_A(\mathrm{GD})$. $\square$

Next we give a decidable characterisation of GD-measurability for regular languages. The characterisation is not so much simple as the one of RD-measurability stated in Theorem 1, but the proof is a natural generalisation of the proof of Theorem 1.

Theorem 4. *Let $\mathcal{A} = (Q, \cdot, q_0, F)$ be a deterministic automaton and let $Q_1, \cdots, Q_k$ be its all sink components and let $Q' = Q \setminus \bigcup_{i=1}^{k} Q_i$. Define*

$$P_i = \{w \in A^* \mid q_0 \cdot w \in Q_i\} \qquad P' = \{w \in A^* \mid q_0 \cdot w \in Q'\}$$
$$S_i = \{w \in A^* \mid Q_i \cdot w \subseteq F\} \qquad S'_i = \{w \in A^* \mid Q_i \cdot w \subseteq Q \setminus F\}$$

for each $i \in \{1, \cdots, k\}$, and define

$$M = \bigcup_{i=1}^{k} P_i S_i \qquad \text{and} \qquad M' = \bigcup_{i=1}^{k} P_i S'_i.$$

Then $L = L(\mathcal{A})$ is GD*-measurable if and only if $\delta_A(L) = \delta_A(M)$ and $\delta_A(\overline{L}) = \delta_A(M')$ holds.*

Proof. By the construction, clearly $M \subseteq L$ and $M' \subseteq \overline{L}$ holds. Also, by Lemma 3, we have $M = \bigcup_{i=1}^{k} P_i A^* S_i$ and $M' = \bigcup_{i=1}^{k} P_i A^* S'_i$. Intuitively, M and M' are "largest" (with respect to the density) languages of the form PA^*S included in

L and $\overline{L}$, respectively. "if" part is easy. $\delta_A(L) = \delta_A(M)$ and $\delta_A(\overline{L}) = \delta_A(M')$ implies that the two sequences of generalised definite languages $M_n = \bigcup_{i=1}^{k}\{uA^*v \mid u \in P_i, v \in S_i, |u| + |v| \leq n\}$ and the complements of $M'_n = \bigcup_{i=1}^{k}\{uA^*v \mid u \in P_i, v \in S'_i, |u| + |v| \leq n\}$ converges to L if n tends to infinity from inner and outer, respectively.

Next we show "only if" part by contraposition. With out loss of generality, we can assume that $\delta_A(L) > \delta_A(M)$. For every $u, v \in A^*$, we show that

$$uA^*v \subseteq L \Rightarrow (uA^* \setminus P')v \subseteq M. \qquad (\diamondsuit)$$

This implies $\delta_A(uA^*v) = \delta_A((uA^* \setminus P')v) \leq \delta_A(M)$ (because P' has density zero by Lemma 3), from this we can conclude that every generalised definite language should have density less than or equal to the density of M. Hence, no sequence of generalised definite languages converges to L from inner by the assumption $\delta_A(L) > \delta_A(M)$. Let $u, v \in A^*$ be words satisfying $uA^*v \subseteq L$, and let uw be a word in $uA^* \setminus P'$. Because uw is not in P', uw is in P_j for some $j \in \{1, \cdots, k\}$. The condition $uA^*v \subseteq L$ implies $uwA^*v \subseteq L$ and hence we have $uww'v \in L$ for any word $w' \in A^*$. For every $q \in Q_j$, there is some word w' such that $q_0 \cdot uww' = q$ because Q_j is strongly connected. Thus we can conclude that $q \cdot v \in F$ for each $q \in Q_j$, which means that v is in S_j and hence uwv is in M (by $uw \in P_j$ and $v \in S_j$), *i.e.*, the condition ($\diamondsuit$) is true. Let $R = E \cup \bigcup_{i \in I} F_i A^* G_i$ be a generalised definite language included in L, where E and $F_i, G_i \subseteq A^*$ are finite for all $i \in I$ and I is a finite index set. The condition ($\diamondsuit$) and $R \subseteq L$ implies that $\bigcup_{i \in I}(F_i A^* \setminus P')G_i \subseteq M$ (note that E is density zero because it is finite). This means that any generalised definite subset of L should have a density smaller or equal to $\delta_A(M)$ which is strictly smaller than $\delta_A(L)$ by the assumption. Thus there is no convergent sequence of generalised definite languages to L from inner. $\square$

By the construction, clearly, all languages P_i, S_i, S'_i are regular and automata recognising these languages can be constructed from $\mathcal{A}$. Hence, we can effectively construct two automata recognising M and M' from $\mathcal{A}$. Also, checking the condition $\delta_A(L) = \delta_A(M)$ and $\delta_A(\overline{L}) = \delta_A(M')$ is decidable: this condition is equivalent to $\delta_A(M \cup M') = 1$, and it is decidable in linear time whether a given deterministic automaton recognises a co-null regular language (*cf.* [13]).

Corollary 2. *For a given regular language L it is decidable whether L is* GD-*measurable (equivalently,* LT-*measurable by Proposition 1).*

4.1 Remark on the measuring power of GD

As we stated in Example 2, the language $M_k = \{w \in \{a, b\}^* \mid |w|_a = |w|_b \mod k\}$ is LT-*im*measurable for any $k \geq 2$. The proof of the above fact given in [16] uses an algebraic characterisation of locally testable languages. However, through Proposition 1, we can more easily prove this fact by showing that M_k is LT-immeasurable as follows.

Proposition 2. $M_k = \{w \in \{a,b\}^* \mid |w|_a = |w|_b \mod k\}$ *is* GD*-immeasurable for any* $k \geq 2$.

Proof. By simple calculation, we have $\delta_A(M_k) = 1/k$. By definition, every infinite generalised definite language must contain a language of the form uA^*v for some $u, v \in A^*$. Let $n = |uv|_a - |uv|_b \mod k$, define $w = b$ if $n = 0$ and $w = \varepsilon$ otherwise. Then we have $uwv \in L$ but $uwv \notin M_k$. This means that $\underline{\mu}_{\mathrm{GD}}(M_k) = 0 < \delta_A(M_k)$, *i.e.*, M_k is GD-immeasurable. □

A non-empty word w is said to be *primitive* if there is no shorter word v such that $w = v^k$ for some $k \geq 2$. In [14], it is shown that the set Q of all primitive words over $A = \{a, b\}$ is REG-immeasurable where REG is the class of all regular languages. The proof given in [14] involves some non-trivial analysis of the syntactic monoid of a regular language. If we consider the more weaker notion, GD-measurability, the proof of the GD-immeasurability is almost trivial: by definition, every infinite generalised definite language must contain a language of the form uA^*v. But uA^*v contains the non-primitive word $uvuv$, hence there is no infinite generalised definite subset of Q.

From the last example, one can naturally consider that the GD-measurability is a very weaker notion than the REG-measurability. We are interested in how far the GD-measurability is from the REG-measurability: is there any natural subclass $\mathrm{GD} \subsetneq \mathcal{C} \subsetneq \mathrm{REG}$ of regular languages such that the $\mathcal{C}$-measurability differs from these two measurability? A possible candidate is SF the class of all star-free languages as we discussed in the next section.

5 Related and Future Work

As we stated in Section 1, the decidability of SF-measurability [15] for regular languages is still unknown. The decidability of LT-measurability was left open in [16], but thanks to Proposition 1 and Theorem 4, it was shown that LT-measurability (= GD-measurability) is decidable.

For some weaker fragments of star-free languages, the decidability of measurability for regular languages are known: a language L is called *piecewise testable* [12] if it can be represented as a finite Boolean combination of languages of the form $A^*a_1A^* \cdots A^*a_kA^*$ (where $a_i \in A$ for each i), and L is called *alphabet testable* if it can be represented as a finite Boolean combination of languages of the form A^*aA^* (where $a \in A$). We denote by PT and AT the class of all piecewise testable and alphabet testable languages, respectively. It was shown in [16] that AT-measurability and PT-measurability are both decidable. Moreover, AT-measurability and PT-measurability do not rely on the existence of an infinite convergent sequence, but rely on the existence of a certain *single* language [16]:

- L is AT-measurable if and only if L or its complement contains $\bigcap_{a \in A} A^*aA^*$.
- L is PT-measurable if and only if L or its complement contains a language of the form $A^*a_1A^* \cdots A^*a_kA^*$

In [17] the tight complexity bounds of AT-measurability and PT-measurability for regular languages was given: AT-measurability is *co*-**NP**-*complete* and PT-measurability is *decidable in linear time*, if an input regular language is given by a deterministic automaton. Even though AT is a very restricted subclass of PT, the complexity of AT-measurability is much higher than PT-measurability. This contrast is interesting. Thanks to Theorem 1, RD-measurability is decidable in linear time, if an input regular language is given by a minimal automaton.

Our future work are three kinds.

(1) Give the tight complexity bound of D- and GD-measurability.
(2) Prove or disprove $\mathrm{Ext}_A(\mathrm{GD}) \subsetneq \mathrm{Ext}_A(\mathrm{SF})$.
(3) If $\mathrm{Ext}_A(\mathrm{GD}) \subsetneq \mathrm{Ext}_A(\mathrm{SF})$, prove or disprove the decidability of SF-measurability.

As demonstrated in the proof of Theorem 4, GD-measurability heavily relies on the existence of an *infinite sequence* of different generalised definite languages. Hence the situation is essentially different with AT-measurability and PT-measurability. One might naturally consider that GD-measurability has a more higher complexity than AT-measurability.

To tackle the problem (2) and (3), perhaps we can use some known techniques related to star-free languages, for example, the so-called *separation problem* for a language class $\mathcal{C}$: for a given pair of regular languages (L_1, L_2), is there a language L in $\mathcal{C}$ such that $L_1 \subseteq L$ and $L \cap L_2 = \emptyset$ (L "separates" L_1 and L_2)? It is known that the separation problem for SF is decidable [10].

Acknowledgements: I am grateful to Mark V. Lawson whose helpful discussions were extremely valuable. The author also thank to anonymous reviewers for many valuable comments. This work was supported by JST ACT-X Grant Number JPMJAX210B, Japan.

References

1. 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. pp. 366–380 (2014)
2. Berstel, J., Perrin, D., Reutenauer, C.: Codes and Automata, Encyclopedia of mathematics and its applications, vol. 129. Cambridge University Press (2010)
3. Brzozowski, J.: Canonical regular expressions and minimal state graphs for definite events. Mathematical Theory of Automata **12**, 529–561 (1962)
4. Brzozowski, J.: On Aperiodic I-monoids. Department of Computer Science. University of Waterloo. Research Report CS-75-28 (1975)
5. Brzozowski, J., Simon, I.: Characterizations of locally testable events. vol. 4, pp. 243–271 (1973)
6. Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press, New York, NY, USA, 1 edn. (2009)
7. Ginzburg, A.: About some properties of definite, reverse-definite and related automata. IEEE Transactions on Electronic Computers **EC-15**(5), 806–810 (1966)
8. Lawson, M.V.: Finite Automata. Chapman and Hall/CRC (2004)
9. McNaughton, R.: Algebraic decision procedures for local testability. Mathematical systems theory **8**, 60–76 (1974)
10. Place, T., Zeitoun, M.: Separating regular languages with first-order logic. Log. Methods Comput. Sci. **12**(1) (2016)
11. Salomaa, A., Soittola, M.: Automata Theoretic Aspects of Formal Power Series. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1978)
12. Simon, I.: Piecewise testable events. In: Automata Theory and Formal Languages. pp. 214–222 (1975)
13. Sin'ya, R.: An automata theoretic approach to the zero-one law for regular languages. In: Games, Automata, Logics and Formal Verification. pp. 172–185 (2015)
14. Sin'ya, R.: Asymptotic approximation by regular languages. In: Current Trends in Theory and Practice of Computer Science. pp. 74–88 (2021)
15. Sin'ya, R.: Carathéodory extensions of subclasses of regular languages. In: Developments in Language Theory. pp. 355–367 (2021)
16. Sin'ya, R.: Measuring power of locally testable languages. In: Diekert, V., Volkov, M. (eds.) Developments in Language Theory. pp. 274–285. Springer International Publishing, Cham (2022)
17. Sin'ya, R., Yamaguchi, Y., Nakamura, Y.: Regular languages that can be approximated by testing subword occurrences. Computer Software **40**(2), 49–60 (2023), (written in Japanese)
18. Zalcstein, Y.: Locally testable languages. Journal of Computer and System Sciences **6**(2), 151–167 (1972)