On the Average State Complexity of Shuffle Ideals

Guilherme Duarte[0000−0002−4119−0694]1, Markus Holzer[0000−0003−4224−4014]2, Nelma Moreira[0000−0003−0861−0105]1, and Rogério Reis[0000−0001−9668−0917]1 ⋆

[1] CMUP & DCC, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal.
{guilherme.duarte,nelma.moreira,rogerio.reis}@fc.up.pt
[2] Institut für Informatik, Universität Giessen, Arndtstr. 2, 35392 Giessen, Germany.
holzer@informatik.uni-giessen.de

Abstract. We consider shuffle ideals and their finite bases, both represented by regular expressions. We introduce an operation that, given a regular expression $\alpha \in \mathsf{RE}_{\text{fin}}$ representing a finite language, returns an expression for the shuffle ideal $\mathcal{L}(\alpha) \sqcup\!\sqcup\, \Sigma^\star$. This operation induces a new combinatorial class, which we call RE_{SI}. For both classes, we define the Antimirov automaton and provide results on their state complexity in the worst as well as the average case. For $\alpha \in \mathsf{RE}_{\text{fin}}$, in the worst case, the number of states is at most $|\alpha|_\Sigma$, the number of letters in α. The asymptotic average is, however, bounded by $\frac{|\alpha|_\Sigma}{2}$. For $\tau \in \mathsf{RE}_{\text{SI}}$, in the worst case, the number of states is at most $|\tau|_{\Sigma^\star}$, the number of occurrences of $\Sigma^\star$ in τ. On the other hand, the asymptotic average number of states is bounded by $\frac{|\tau|_{\Sigma^\star}}{2}$.

1 Introduction

The family of shuffle ideals plays a key role in formal language theory, connecting combinatorics on words, automata theory, the algebraic study of languages, and applications, see, e.g., [1,10,14,15] to mention a few. A shuffle ideal is a language that is closed under the shuffle operation. This means that if a word belongs to the language, then so does every word that is formed by interleaving it with an arbitrary word from the same alphabet. Alternatively, shuffle ideals are upward-closed under the subsequence order: if a language L defined over an alphabet is a shuffle ideal and u is a word in L that is a subsequence of v, then v is also in L. In concrete terms, shuffle ideals are precisely the languages of the form

$$\Sigma^\star a_1 \Sigma^\star a_2 \Sigma^\star \cdots \Sigma^\star a_n \Sigma^\star \qquad \text{where } a_1, a_2, \ldots, a_n \in \Sigma, \text{ for some } n \geq 0,$$

and their finite unions—languages defined purely by the presence of a fixed subsequence pattern, independent of surrounding context that form the $\frac{1}{2}$-level

⋆ Research partially supported by CMUP through FCT under the project with DOI 10.54499/UID/00144/2025 and by FCT project AVATAR 2023.12169.PEX. Guilherme Duarte is also supported by the FCT grant 2024.01528.BD.

of the Straubing-Thérien Hierarchy (STH) [16, 17]. The combination of their combinatorial richness, algebraic structure, and closure under interleaving makes shuffle ideals both theoretically intriguing and practically relevant.

Mostly, the descriptional complexity of shuffle ideals has been studied extensively, particularly in terms of finite automata and the structure of their generating sets. Some work focuses on characterising shuffle ideals *via* finite automaton representations, analysing the size and structure of minimal deterministic and nondeterministic automata accepting such languages [8–10]. In this context, several results establish upper and lower bounds on state complexity, highlighting the potential blow-up caused by closure under interleaving. Another line of research investigates the notion of a finite word basis, where a shuffle ideal is represented as the upward closure of a finite set of words under the subsequence order [11]. These studies examine the size and minimality of such bases, as well as their relationship to automata representations [3, 12]. Collectively, prior work provides a detailed understanding of the trade-offs between automata size, structural properties of shuffle ideals, and the complexity of their finite generators. To our knowledge the relation between regular expressions describing shuffle ideals and finite automata size was not studied so far. This is the starting point for our investigation.

We study the state complexity of Antimirov automata [2] (also known as partial derivative automata) for two specifically defined combinatorial classes: $\mathsf{RE}_{\mathrm{fin}}$, describing finite languages, and $\mathsf{RE}_{\mathrm{SI}}$, representing exactly shuffle ideals. Given $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, we introduce an operation $\mathsf{e}(\alpha)$ that returns an expression that generates the shuffle ideal $\mathcal{L}(\alpha) \sqcup\!\sqcup \Sigma^\star$, nicely relating $\mathsf{RE}_{\mathrm{fin}}$ with $\mathsf{RE}_{\mathrm{SI}}$. To both of these classes, we define the support of an expression, that corresponds to the set of states of the Antimirov automaton and is suitable for obtaining average complexity estimates. We show that the worst-case state complexity of Antimirov automaton for $\mathsf{RE}_{\mathrm{fin}}$ is bounded by the number of alphabet symbols in the expression, denoted by $|\alpha|_\Sigma$, while for $\mathsf{RE}_{\mathrm{SI}}$, it is limited by the number of occurrences of $\Sigma^\star$, denoted as $|\tau|_{\Sigma^\star}$ and counting $\Sigma^\star$ a constant operator in the expression. Counting $\Sigma^\star$ as a "single symbol" in the expression is reasonable in this context, since in practical regular expressions it is often written as `.*`, e.g., in Unix/Linux regular expressions, which conceptually represents a single wildcard for arbitrary-length strings. Utilizing the framework of analytic combinatorics [7] and generating functions, we conclude with the finding that the asymptotic average number of states is bounded by $\frac{|\alpha|_\Sigma}{2}$ for $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, and by $\frac{|\tau|_{\Sigma^\star}}{2}$ for $\tau \in \mathsf{RE}_{\mathrm{SI}}$. By comparing both results we obtain that the partial derivate automaton for $\mathsf{e}(\alpha) \in \mathsf{RE}_{\mathrm{SI}}$, with $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, is on average 2.45 times larger than the size of the corresponding partial derivative automaton for α.

2 Preliminaries

Given a set S, we denote its cardinality by $|S|$ and 2^S the collection of all its subsets. An *alphabet* is a finite set of *letters* (or alphabet symbols) $\Sigma = \{a_1, a_2, \ldots, a_k\}$. A *word* $w = b_1 b_2 \cdots b_n$ is a sequence of symbols, where each

$b_i \in \Sigma$. The length of a word w is denoted by $|w|$, and we denote by ε the empty-word. A *language* is a set of words over an alphabet, Σ^+ denotes the language containing all non-empty words over Σ, and $\Sigma^\star = \Sigma^+ \cup \{\varepsilon\}$. The *left quotient* of a language L w.r.t. a letter $a \in \Sigma$ is $a^{-1}L = \{\, w \mid aw \in L \,\}$. The left quotient of a language L w.r.t. a word $w \in \Sigma^\star$ is then inductively defined by $\varepsilon^{-1}L = L$ and $(wa)^{-1}L = a^{-1}(w^{-1}L)$.

A *nondeterministic finite automaton* (NFA) is a tuple $\mathcal{A} = \langle Q, \Sigma, q_0, \delta, F\rangle$, where Q is a finite set of states, $\delta : Q \times \Sigma \to 2^Q$ is the transition function, $q_0 \in Q$ is the initial state, and $F \subseteq Q$ is the set of final states. The language recognised by $\mathcal{A}$ is denoted by $\mathcal{L}(\mathcal{A})$. In this paper, we consider the *size* of an NFA as $|Q|$.

2.1 Shuffle and Shuffle Ideals

Given an alphabet Σ, the *shuffle* of two words $u, v \in \Sigma^\star$ is a finite set of words inductively defined as follows:

$$\begin{aligned} u \sqcup\!\sqcup\, \varepsilon &= \varepsilon \sqcup\!\sqcup\, u = \{u\}, \\ au \sqcup\!\sqcup\, bv &= \{\, aw \mid w \in u \sqcup\!\sqcup\, bv \,\} \cup \{\, bw \mid w \in au \sqcup\!\sqcup\, v \,\}, \end{aligned}$$

where $a, b \in \Sigma$. This definition is extended to languages $L, L' \subseteq \Sigma^\star$ in the natural way:

$$L \sqcup\!\sqcup\, L' = \bigcup_{u \in L, v \in L'} u \sqcup\!\sqcup\, v.$$

It is well-known that regular languages are closed under the shuffle operation.

A *shuffle ideal* is a language which is a finite union of languages of the form $\Sigma^\star a_1 \Sigma^\star a_2 \Sigma^\star \cdots \Sigma^\star a_n \Sigma^\star$, where $a_1, a_2, \ldots, a_n \in \Sigma$, for some $n \geq 0$ (possibly different for each language in the union).

Theorem 1 (Higman [11]). *A language L is a shuffle ideal if and only if there exists a finite language K such that $L = K \sqcup\!\sqcup\, \Sigma^\star$. The language K is called a* basis *of the shuffle ideal L.*

If a shuffle ideal L is represented as above the set of the words $a_1 a_2 \cdots a_n$ is a basis of L.

Example 1. Let $\Sigma = \{a, b\}$. The language $\Sigma^\star$ is a shuffle ideal with basis $\{\varepsilon\}$. The language $\{\, w \in \Sigma^\star \mid a \text{ and } b \text{ occur in } w \,\}$ is a shuffle ideal with basis $\{ab, ba\}$. The language $L = \mathcal{L}((aa + bb + (ab + ba)(a + b))(a + b)^\star)$ is a shuffle ideal with basis $K = \{aa, bb, baa\}$.

The following lemmas are immediate consequences of the definition of shuffle and shuffle ideals, and will be useful in the next sections.

Lemma 1. *Let $L \subseteq \Sigma^\star$. Then, $(L \sqcup\!\sqcup\, \Sigma^\star) \cdot \Sigma^\star = \Sigma^\star \cdot (L \sqcup\!\sqcup\, \Sigma^\star) = L \sqcup\!\sqcup\, \Sigma^\star$.*

Lemma 2. *Let $L, L' \subseteq \Sigma^\star$. Then, $LL' \sqcup\!\sqcup\, \Sigma^\star = (L \sqcup\!\sqcup\, \Sigma^\star)(L' \sqcup\!\sqcup\, \Sigma^\star)$.*

3 Regular Expressions for Finite Languages and Shuffle Ideals

In this section, we consider regular expressions (REs) for finite languages. We denote this class of REs by RE_{fin} containing all the expressions α generated by the grammar

$$\alpha \to \emptyset \mid \varepsilon \mid \alpha_{\varepsilon} \mid \alpha_{\overline{\varepsilon}} \tag{1}$$

$$\alpha_{\varepsilon} \to \alpha_{\overline{\varepsilon}}^{?} \mid (\alpha_{\varepsilon} + \alpha_{\varepsilon}) \mid (\alpha_{\varepsilon} + \alpha_{\overline{\varepsilon}}) \mid (\alpha_{\overline{\varepsilon}} + \alpha_{\varepsilon}) \mid (\alpha_{\varepsilon} \cdot \alpha_{\varepsilon}) \tag{2}$$

$$\alpha_{\overline{\varepsilon}} \to a \mid (\alpha_{\overline{\varepsilon}} + \alpha_{\overline{\varepsilon}}) \mid (\alpha_{\overline{\varepsilon}} \cdot \alpha_{\overline{\varepsilon}}) \mid (\alpha_{\overline{\varepsilon}} \cdot \alpha_{\varepsilon}) \mid (\alpha_{\varepsilon} \cdot \alpha_{\overline{\varepsilon}}) \qquad (a \in \Sigma), \tag{3}$$

over an alphabet Σ, where the $\cdot$ and the parentheses are often omitted. As usual, the regular language $\mathcal{L}(\alpha)$ represented by an expression $\alpha \in \mathsf{RE}_{\text{fin}}$ is inductively defined as $\mathcal{L}(\emptyset) = \emptyset$, $\mathcal{L}(\varepsilon) = \{\varepsilon\}$, $\mathcal{L}(a) = \{a\}$, and

$$\mathcal{L}(\alpha_{\overline{\varepsilon}}^{?}) = \mathcal{L}(\alpha_{\overline{\varepsilon}}) \cup \{\varepsilon\}, \;\; \mathcal{L}(\alpha \cdot \beta) = \mathcal{L}(\alpha)\mathcal{L}(\beta), \;\; \mathcal{L}(\alpha + \beta) = \mathcal{L}(\alpha) \cup \mathcal{L}(\beta),$$

where $a \in \Sigma$ and, given $S, T \subseteq \Sigma^{\star}$, $S \cdot T = \{\, xy \mid x \in S, y \in T \,\}$. We say that two expressions $\alpha, \beta \in \mathsf{RE}_{\text{fin}}$ are equivalent, denoted as $\alpha \doteq \beta$, if $\mathcal{L}(\alpha) = \mathcal{L}(\beta)$. Moreover, we say that $\varepsilon(\alpha) = \varepsilon$ if $\varepsilon \in \mathcal{L}(\alpha)$, and $\varepsilon(\alpha) = \emptyset$ otherwise. The expressions α_{ε} and $\alpha_{\overline{\varepsilon}}$ generate all terms $\alpha \in \mathsf{RE}_{\text{fin}}$ such that $\varepsilon(\alpha) = \varepsilon$ (with the exception of the expression ε itself) and $\varepsilon(\alpha) = \emptyset$, respectively.

For the *size* of a regular expression $\alpha \in \mathsf{RE}_{\text{fin}}$, denoted by $\|\alpha\|$, we consider the number of symbols in α excluding parentheses, and is inductively defined as:

$$\|\emptyset\| = \|\varepsilon\| = \|a\| = 1, \;\; \|\alpha_{\overline{\varepsilon}}^{?}\| = \|\alpha_{\overline{\varepsilon}}\| + 1, \;\; \|\alpha \cdot \beta\| = \|\alpha + \beta\| = \|\alpha\| + 1 + \|\beta\|.$$

Here, both α and β can be either an expression of the form α_{ε} or $\alpha_{\overline{\varepsilon}}$.

Now, we shall consider regular expressions for shuffle ideals. We denote this class of REs by RE_{SI}, containing all the expressions τ generated by the grammar

$$\tau \to \emptyset \mid \Sigma^{\star} \mid \tau_{\mathtt{LR}} \tag{4}$$

$$\tau_{\mathtt{LR}} \to \Sigma^{\star}\, a\, \Sigma^{\star} \mid (\tau_{\mathtt{LR}} + \tau_{\mathtt{LR}}) \mid (\tau_{\mathtt{LR}} \cdot \tau_{\mathtt{R}}) \tag{5}$$

$$\tau_{\mathtt{R}} \to a\, \Sigma^{\star} \mid (\tau_{\mathtt{R}} + \tau_{\mathtt{R}}) \mid (\tau_{\mathtt{R}} \cdot \tau_{\mathtt{R}}), \tag{6}$$

for $a \in \Sigma$. Here, $\Sigma^{\star}$ [3] is a terminal symbol from the grammar that represents the language $\Sigma^{\star}$, i.e., $\mathcal{L}(\Sigma^{\star}) = \Sigma^{\star}$. The regular language $\mathcal{L}(\tau)$ represented by an expression $\tau \in \mathsf{RE}_{\text{SI}}$ follows the same inductive definition as for $\alpha \in \mathsf{RE}_{\text{fin}}$. For the size of τ, we use

$$\|\, \Sigma^{\star} \,\| = 1, \;\; \|\, \Sigma^{\star}\, a\, \Sigma^{\star} \,\| = 3, \;\; \|a\, \Sigma^{\star} \,\| = 2,$$

for $a \in \Sigma$, and the inductive cases are the same as for expressions in RE_{fin}. The function $\varepsilon(\cdot)$ and the relation $\doteq$ also follow analogously for expressions in RE_{SI}.

[3] We follow the convention of using the symbol $\Sigma^{\star}$ to denote the language $\Sigma^{\star}$ in the context of shuffle ideals.

Before showing that $\mathsf{RE}_{\mathrm{SI}}$ represents shuffle ideals, we build a bridge between the classes $\mathsf{RE}_{\mathrm{fin}}$ and $\mathsf{RE}_{\mathrm{SI}}$, by defining an operation $\mathsf{e} : \mathsf{RE}_{\mathrm{fin}} \to \mathsf{RE}_{\mathrm{SI}}$. Informally, given $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, the expression $\mathsf{e}(\alpha)$ is given by inserting $\Sigma^\star$ around every alphabet symbol in α.

Definition 1. *Let $\alpha \in \mathsf{RE}_{\mathrm{fin}}$ over an alphabet Σ. Then, $\mathsf{e}(\alpha)$ is inductively defined as*

$$\begin{aligned} \mathsf{e}(\emptyset) &= \emptyset, & \mathsf{e}(\varepsilon) &= \mathsf{e}(\alpha_\varepsilon) = \Sigma^\star\,, \\ \mathsf{e}(a) &= \Sigma^\star\, a\, \Sigma^\star\,, & \mathsf{e}(\alpha_{\overline{\varepsilon}} + \beta_{\overline{\varepsilon}}) &= \mathsf{e}(\alpha_{\overline{\varepsilon}}) + \mathsf{e}(\beta_{\overline{\varepsilon}}), \\ \mathsf{e}(\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) = \mathsf{e}(\alpha_{\overline{\varepsilon}} \cdot \alpha_\varepsilon) &= \mathsf{e}(\alpha_{\overline{\varepsilon}}), & \mathsf{e}(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) &= \mathsf{e}(\alpha_{\overline{\varepsilon}}) \cdot \overrightarrow{\mathsf{e}}(\beta_{\overline{\varepsilon}}), \end{aligned}$$

where

$$\begin{aligned} \overrightarrow{\mathsf{e}}(a) &= a\, \Sigma^\star\,, & \overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}} + \beta_{\overline{\varepsilon}}) &= \overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}}) + \overrightarrow{\mathsf{e}}(\beta_{\overline{\varepsilon}}), \\ \overrightarrow{\mathsf{e}}(\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) = \overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}} \cdot \alpha_\varepsilon) &= \overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}}), & \overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) &= \overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}}) \cdot \overrightarrow{\mathsf{e}}(\beta_{\overline{\varepsilon}}), \end{aligned}$$

for $a \in \Sigma$.

We emphasize that $\mathsf{e}(\alpha_\varepsilon) = \Sigma^\star$, that is, any expression in $\mathsf{RE}_{\mathrm{fin}}$ that generates a language containing ε is mapped to the same expression $\Sigma^\star$ by this operation.

The use of the auxiliary operation $\overrightarrow{\mathsf{e}}(\cdot)$ avoids the concatenation of two $\Sigma^\star$-expressions in the resulting expression $\mathsf{e}(\alpha)$. In fact, we have

Lemma 3. *Let $\alpha_{\overline{\varepsilon}}, \beta_{\overline{\varepsilon}} \in \mathsf{RE}_{\mathrm{fin}}$. Then, $\mathsf{e}(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) \doteq \mathsf{e}(\alpha_{\overline{\varepsilon}}) \cdot \mathsf{e}(\beta_{\overline{\varepsilon}})$.*

Now, we shall prove that $\mathsf{e}(\alpha)$ is an expression for the shuffle ideal $\mathcal{L}(\alpha) \sqcup\!\sqcup\, \Sigma^\star$. First, notice that $\mathcal{L}(\varepsilon) \sqcup\!\sqcup\, \Sigma^\star = \Sigma^\star$. In fact, the same holds for any expression $\alpha_\varepsilon \in \mathsf{RE}_{\mathrm{fin}}$, as the following lemma states.

Lemma 4. *Let $\alpha_\varepsilon \in \mathsf{RE}_{\mathrm{fin}}$. Then $\mathcal{L}(\alpha_\varepsilon) \sqcup\!\sqcup\, \Sigma^\star = \Sigma^\star$.*

Proposition 1. *For $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, $\mathcal{L}(\mathsf{e}(\alpha)) = \mathcal{L}(\alpha) \sqcup\!\sqcup\, \Sigma^\star$.*

Proof. The proof follows by induction on the structure of α. If $\alpha = \emptyset$, $\alpha = \varepsilon$, or $\alpha = a$, with $a \in \Sigma$, the result is immediate. Moreover, by Lemma 4, we have that $\mathcal{L}(\mathsf{e}(\alpha_\varepsilon)) = \mathcal{L}(\alpha_\varepsilon) \sqcup\!\sqcup\, \Sigma^\star = \Sigma^\star$.

Now, suppose that the claim is true for $\alpha_{\overline{\varepsilon}}$. Three cases arise:

- For $\alpha = \alpha_{\overline{\varepsilon}} + \beta_{\overline{\varepsilon}}$, we have

  $$\mathcal{L}(\alpha) \sqcup\!\sqcup\, \Sigma^\star = (\mathcal{L}(\alpha_{\overline{\varepsilon}}) + \mathcal{L}(\beta_{\overline{\varepsilon}})) \sqcup\!\sqcup\, \Sigma^\star = \mathcal{L}(\alpha_{\overline{\varepsilon}}) \sqcup\!\sqcup\, \Sigma^\star \cup \mathcal{L}(\beta_{\overline{\varepsilon}}) \sqcup\!\sqcup\, \Sigma^\star\,,$$

  thus the result follows;
- For $\alpha = \alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}$, we have

  $$\mathcal{L}(\alpha) \sqcup\!\sqcup\, \Sigma^\star = (\mathcal{L}(\alpha_\varepsilon)\mathcal{L}(\alpha_{\overline{\varepsilon}})) \sqcup\!\sqcup\, \Sigma^\star = (\mathcal{L}(\alpha_\varepsilon) \sqcup\!\sqcup\, \Sigma^\star)(\mathcal{L}(\alpha_{\overline{\varepsilon}}) \sqcup\!\sqcup\, \Sigma^\star),$$

  from Lemma 2. It follows that $\mathcal{L}(\alpha_\varepsilon) \sqcup\!\sqcup\, \Sigma^\star = \Sigma^\star$, by Lemma 4, and thus $\mathcal{L}(\alpha) \sqcup\!\sqcup\, \Sigma^\star = \mathcal{L}(\alpha_{\overline{\varepsilon}}) \sqcup\!\sqcup\, \Sigma^\star$, from Lemma 1. The proof follows analogously for $\alpha = \alpha_{\overline{\varepsilon}} \cdot \alpha_\varepsilon$;

- For $\alpha = \alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}$, we have

$$\mathcal{L}(\alpha) \sqcup\!\sqcup \Sigma^\star = (\mathcal{L}(\alpha_{\overline{\varepsilon}}) \cdot \mathcal{L}(\beta_{\overline{\varepsilon}})) \sqcup\!\sqcup \Sigma^\star = (\mathcal{L}(\alpha_{\overline{\varepsilon}}) \sqcup\!\sqcup \Sigma^\star)(\mathcal{L}(\beta_{\overline{\varepsilon}}) \sqcup\!\sqcup \Sigma^\star),$$

by Lemma 2. The result follows by the induction hypothesis and Lemma 3. □

Proposition 2. *The class* $\mathsf{RE}_{\mathrm{SI}}$ *is equal to the class* $\{\, \mathsf{e}(\alpha) \mid \alpha \in \mathsf{RE}_{\mathrm{fin}} \,\}$.

The following lemma gives an upper bound on the size of $\mathsf{e}(\alpha)$, as a function of the size of $\alpha \in \mathsf{RE}_{\mathrm{fin}}$. This will be useful in the next sections, when we analyse the state complexity of the Antimirov automaton for $\mathsf{e}(\alpha)$.

Lemma 5. *Let* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$. *Then,* $\|\mathsf{e}(\alpha)\| \leq 2\|\alpha\| + 1$.

4 Support for a Regular Expression

A *support* for a regular expression is a finite set of regular expressions that captures all possible *residuals* of the expression after reading one initial letter [13]. Thus, a support gives a finite recursive description of the language of the expression, and gives rise to an NFA that accepts the same language. Consider the alphabet $\Sigma = \{a_1, a_2, \ldots, a_k\}$ and an expression $\alpha_0 \in \mathsf{RE}_{\mathrm{fin}}$. Formally, a support of α_0 is a set $\{\alpha_1, \alpha_2, \ldots, \alpha_n\} \subseteq \mathsf{RE}_{\mathrm{fin}}$ that satisfies the system of equations

$$\alpha_i \doteq a_1\alpha_{i,1} + a_2\alpha_{i,2} + \ldots + a_k\alpha_{i,k} + \varepsilon(\alpha_i),$$

for $i \in \{0, 1, \ldots, n\}$ and for some $\alpha_{i,1}, \alpha_{i,2}, \ldots, \alpha_{i,k}$, where each one is a (possible empty) sum of elements in $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$. The existence of a support for a regular expression α implies the existence of an NFA that accepts $\mathcal{L}(\alpha)$, namely, $\mathcal{A} = \langle \{\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n\}, \Sigma, \{\alpha_0\}, \delta, F\rangle$, with $F = \{\, \alpha_i \mid \varepsilon(\alpha_i) = \varepsilon \,\}$ and $\delta(\alpha_i, a_j) = \{\alpha_{i_1}, \alpha_{i_2}, \ldots, \alpha_{i_m}\}$, where $\alpha_{i,j} = \alpha_{i_1} + \alpha_{i_2} + \ldots + \alpha_{i_m}$, for some $m \in \{0, 1, \ldots, n\}$ and indices $i_1, i_2, \ldots, i_m \in \{0, 1, \ldots, n\}$. A support for an expression $\alpha \in \mathsf{RE}_{\mathrm{fin}}$ can be computed using the function $\pi : \mathsf{RE}_{\mathrm{fin}} \to 2^{\mathsf{RE}_{\mathrm{fin}}}$ defined as follows. We do not distinguish between all possible scenarios for union and concatenation, as they are all covered by the same definition.

Definition 2. *Given* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, *let* $\pi(\alpha)$ *be inductively defined as* $\pi(\emptyset) = \pi(\varepsilon) = \emptyset$ *and*

$$\begin{aligned} \pi(a) &= \{\varepsilon\}, & \pi(\alpha^?_{\overline{\varepsilon}}) &= \pi(\alpha_{\overline{\varepsilon}}), \\ \pi(\alpha + \beta) &= \pi(\alpha) \cup \pi(\beta), & \pi(\alpha \cdot \beta) &= \pi(\alpha) \cdot \beta \cup \pi(\beta), \end{aligned}$$

where $a \in \Sigma$ *and, given* $S \subseteq \mathsf{RE}_{\mathrm{fin}}$, $\emptyset \cdot S = S \cdot \emptyset = \emptyset$, $S \cdot \varepsilon = S$ *and for* $\beta \neq \varepsilon$, *if* $\varepsilon \notin S$, $S \cdot \beta = \{\, \alpha \cdot \beta \mid \alpha \in S \,\}$, *otherwise* $S \cdot \beta = \{\, \alpha \cdot \beta \mid \alpha \in S \setminus \{\varepsilon\} \,\} \cup \{\beta\}$.

Proposition 3. *For* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, *the set* $\pi(\alpha)$ *is a support for* α.

Lemma 6. *For* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, $|\pi(\alpha)| \leq |\alpha|_\Sigma$, *where* $|\alpha|_\Sigma$ *denotes the number of letters in* α.

This upper bound is reached if, in every step of the construction of $\pi(\alpha)$, all unions are disjoint. There are, however, cases where $\pi(\alpha) < |\alpha|_\Sigma$ [4].

Lemma 7. *Let* $\alpha \in \mathsf{RE}_{\text{fin}} \setminus \{\varepsilon, \emptyset\}$. *Then,* $\varepsilon \in \pi(\alpha)$.

By the lemma above, we have that

$$|\pi(\alpha + \beta)| = |\pi(\alpha) \cup \pi(\beta)| \leq |\pi(\alpha)| + |\pi(\beta)| - 1. \tag{7}$$

We now define the π-function for $\tau \in \mathsf{RE}_{\text{SI}}$. We also need to provide a definition for the support of an expression generated by $\tau_{\mathtt{R}}$. As before, we do not distinguish between all possible cases of union and concatenation, as they are all covered by the same definition.

Definition 3. *Given* $\tau \in \mathsf{RE}_{\text{SI}}$, *let* $\pi(\tau)$ *be inductively defined as* $\pi(\emptyset) = \emptyset$ *and*

$$\begin{aligned} \pi(\Sigma^\star) = \pi(a\,\Sigma^\star) &= \{\Sigma^\star\}, & \pi(\Sigma^\star\, a\,\Sigma^\star) &= \{\Sigma^\star, \Sigma^\star\, a\,\Sigma^\star\}, \\ \pi(\tau + \tau') &= \pi(\tau) \cup \pi(\tau'), & \pi(\tau \cdot \tau') & \end{aligned}$$

Proposition 4. *For* $\tau \in \mathsf{RE}_{\text{SI}}$, *the set* $\pi(\tau)$ *is a support of* τ.

Lemma 8. *For* $\tau \in \mathsf{RE}_{\text{SI}}$, $|\pi(\tau)| \leq |\tau|_{\Sigma^\star}$, *where* $|\tau|_{\Sigma^\star}$ *denotes the number of occurrences of* $\Sigma^\star$ *in* τ.

Again, this upper bound is reached if, in every step of the construction of $\pi(\tau)$, *all unions are disjoint.*

Lemma 9. *Let* $\tau \in \mathsf{RE}_{\text{SI}} \setminus \{\emptyset\}$. *Then,* $\Sigma^\star \in \pi(\tau)$.

By the lemma above, we have that

$$|\pi(\tau + \tau')| = |\pi(\tau) \cup \pi(\tau')| \leq |\pi(\tau)| + |\pi(\tau')| - 1. \tag{8}$$

5 Partial Derivative Automata

The automaton obtained from the support of a regular expression can alternatively be obtained from the set of partial derivatives of a regular expression [2].

Definition 4. *The set of partial derivatives of a term* $\alpha \in \mathsf{RE}_{\text{fin}}$ *w.r.t. a letter* $a \in \Sigma$, *denoted by* $\partial_a(\alpha)$, *is inductively defined by*

$$\begin{aligned} \partial_a(\emptyset) = \partial_a(\varepsilon) &= \emptyset & \partial_a(\alpha_{\overline{\varepsilon}}^{?}) &= \partial_a(\alpha_{\overline{\varepsilon}}), \\ \partial_a(b) &= \begin{cases} \{\varepsilon\} & \text{if } a = b; \\ \emptyset & \text{otherwise,} \end{cases} & \partial_a(\alpha + \beta) &= \partial_a(\alpha) \cup \partial_a(\beta), \\ \partial_a(\alpha_\varepsilon \cdot \alpha) &= \partial_a(\alpha_\varepsilon) \cdot \alpha \cup \partial_a(\alpha), & \partial_a(\alpha_{\overline{\varepsilon}} \cdot \alpha) &= \partial_a(\alpha_{\overline{\varepsilon}}) \cdot \alpha, \end{aligned}$$

where $b \in \Sigma$, *and the expressions involving sets are as specified in Definition 2.*

The set of partial derivatives of a term $\alpha \in \mathsf{RE}_{\text{fin}}$ w.r.t. a word $w \in \Sigma^\star$, denoted by $\partial_w(\alpha)$, is inductively defined by $\partial_\varepsilon(\alpha) = \{\alpha\}$ and $\partial_{wa}(\alpha) = \partial_a(\partial_w(\alpha))$, where, given a set $S \subseteq \mathsf{RE}_{\text{fin}}$, $\partial_a(S) = \bigcup_{\alpha \in S} \partial_a(\alpha)$. Now, let $\partial(\alpha)$ denote the set of all partial derivatives of an expression $\alpha \in \mathsf{RE}_{\text{fin}}$, i.e., $\partial(\alpha) = \bigcup_{w \in \Sigma^\star} \partial_w(\alpha)$. Also, let $\partial^+(\alpha) = \bigcup_{w \in \Sigma^+} \partial_w(\alpha)$. From [13], we obtain:

Proposition 5. *The set of partial derivatives of a term $\alpha \in \mathsf{RE}_{\text{fin}}$ w.r.t. to non-null words, denoted by $\partial^+(\alpha)$, is inductively obtained by $\partial^+(\emptyset) = \partial^+(\varepsilon) = \emptyset$ and*

$$\partial^+(a) = \{\varepsilon\}, \qquad \partial^+(\alpha^?_{\overline{\varepsilon}}) = \partial^+(\alpha_{\overline{\varepsilon}}),$$
$$\partial^+(\alpha \cdot \beta) = \partial^+(\alpha) \cdot \beta \cup \partial^+(\beta), \qquad \partial^+(\alpha + \beta) = \partial^+(\alpha) \cup \partial^+(\beta),$$

where $a \in \Sigma$.

Corollary 1. *Given $\alpha \in \mathsf{RE}_{\text{fin}}$, one has $\partial^+(\alpha) = \pi(\alpha)$.*

Consequently, $\partial(\alpha)$ corresponds to the set $\{\alpha\} \cup \pi(\alpha)$. It is well known that the set of partial derivatives of a regular expression gives rise to an equivalent NFA [2, 13].

Lemma 10. *Given $\alpha \in \mathsf{RE}_{\text{fin}}$, the* partial derivative automaton *for α, namely,*

$$\mathcal{A}_{\mathsf{PD}} = \langle \partial(\alpha), \Sigma, \alpha, \delta_\alpha, F_\alpha \rangle,$$

where $\delta_\alpha(\beta, a) = \partial_a(\beta)$, $F_\alpha = \{\, \beta \in \partial(\alpha) \mid \varepsilon \in \mathcal{L}(\beta) \,\}$, satisfies $\mathcal{L}(\mathcal{A}_{\mathsf{PD}}) = \mathcal{L}(\alpha)$.

Now, we shall extend the partial derivatives for expressions $\tau \in \mathsf{RE}_{\text{SI}}$. Again, we also consider the partial derivatives for expressions generated by τ_R.

Definition 5. *The set of partial derivatives of $\tau \in \mathsf{RE}_{\text{SI}}$, w.r.t. a letter $a \in \Sigma$, denoted by $\partial_a(\tau)$, is inductively defined by*

$$\partial_a(\emptyset) = \emptyset, \qquad \partial_a(\Sigma^\star) = \{\Sigma^\star\},$$
$$\partial_a(b\,\Sigma^\star) = \begin{cases} \{\Sigma^\star\} & \text{if } a = b; \\ \emptyset & \text{otherwise,} \end{cases} \qquad \partial_a(\Sigma^\star\, b\,\Sigma^\star) = \begin{cases} \{\Sigma^\star, \Sigma^\star\, b\,\Sigma^\star\} & \text{if } a = b; \\ \{\Sigma^\star\, b\,\Sigma^\star\} & \text{otherwise,} \end{cases}$$
$$\partial_a(\tau + \tau') = \partial_a(\tau) + \partial_a(\tau'), \qquad \partial_a(\tau \cdot \tau') = \partial_a(\tau) \cdot \tau',$$

where $b \in \Sigma$.

The set of partial derivatives of an expression $\tau \in \mathsf{RE}_{\text{SI}}$ w.r.t. a word $w \in \Sigma^\star$, denoted as $\partial_w(\tau)$, is extended as before. Let also $\partial(\tau)$ and $\partial^+(\tau)$ be as before, but now for τ-expressions.

Proposition 6. *The set of partial derivatives of a term $\tau \in \mathsf{RE}_{\text{SI}}$, w.r.t. to words $w \in \Sigma^+$, denoted by $\partial^+(\tau)$, is inductively obtained by $\partial^+(\emptyset) = \emptyset$ and*

$$\partial^+(\Sigma^\star) = \partial^+(a\,\Sigma^\star) = \{\Sigma^\star\}, \qquad \partial^+(\Sigma^\star\, a\,\Sigma^\star) = \{\Sigma^\star, \Sigma^\star\, a\,\Sigma^\star\},$$
$$\partial^+(\tau + \tau') = \partial^+(\tau) \cup \partial^+(\tau'), \qquad \partial^+(\tau \cdot \tau') = \partial^+(\tau) \cdot \tau' \cup \partial^+(\tau'),$$

for $a \in \Sigma$.

Corollary 2. *Given $\tau \in \mathsf{RE}_{\mathrm{SI}}$, one has $\partial^+(\tau) = \pi(\tau)$.*

As before, $\partial(\tau)$ corresponds to the set $\{\tau\} \cup \pi(\tau)$. The partial derivative automata are naturally extended to deal with expressions $\tau \in \mathsf{RE}_{\mathrm{SI}}$.

6 Average State Complexity

We now estimate the asymptotic average number of states in the partial derivative automaton for an expression $\alpha \in \mathsf{RE}_{\mathrm{fin}}$, as well as for an expression $\tau \in \mathsf{RE}_{\mathrm{SI}}$, and provide a comparison between the results for the two classes of expressions. This is carried out using the standard methods of analytic combinatorics, as developed by Flajolet and Sedgewick [7], applied to generating functions.

Let A be a combinatorial class equipped with a size function mapping each object to a non-negative integer. Moreover, given some measure of the objects in A, let a_n be the sum of the values of the measure for all objects in A of size n. Then, the generating function associated with A with the given measure is the function $A(z)$ whose power series expansion satisfies $A(z) = \sum_{n \geq 0} a_n z^n$. We use the notation $[z^n]A(z)$ for the coefficient a_n. Generating functions are complex analytic functions, and their coefficients can be estimated by analysing their behaviour around their dominant singularities. For an introduction to this approach applied to formal languages, we refer to [5]. For a more in-depth analysis, we refer to [6].

Theorem 2 ([7]). *The coefficients of the function $f(z) = (1-z)^{-\alpha}$, where $\alpha \in \mathbb{C} \setminus \mathbb{N}_0^-$, have the following asymptotic behaviour:*

$$[z^n]f(z) = \frac{n^{\alpha-1}}{\Gamma(\alpha)} + o\left(n^{\alpha-1}\right),$$

where Γ is Euler's gamma function.

From an unambiguous specification of a combinatorial class (as, for example, a context-free grammar), one can obtain a system of equations for the generating function $w(z)$ whose coefficients we want to estimate, by using the symbolic method *[7]. An algebraic curve $C(z,w)$ for $w(z)$ can be deduced using Buchberger algorithm to obtain a Gröbner basis for the ideal generated by those equations. We then obtain an algebraic curve $C(z,w)$ in $\mathbb{Q}[z,w]$ such that $C(z,w(z)) = 0$. Since $w(z)$ is the generating function of a combinatorial class, it has non-negative coefficients and, by Pringsheim's Theorem [7, Thm IV.6], its dominant singularity ρ is in $\mathbb{R}^+$, and is smaller than or equal to 1. With this singularity, two cases can arise:*

Case I: $\lim_{z \to p} w(z) = a$, *with $a \in \mathbb{R}^+$;*
Case II: $\lim_{z \to p} w(z) = +\infty$.

In the first case, the singularity ρ is a root of the resultant of $C(z,w)$ and $\frac{\partial C}{\partial w}(z,w)$ with respect to w. Similarly, the constant a is a root of the resultant of

$C(z,w)$ and $\frac{\partial C}{\partial z}(z,w)$ with respect to z. In the second case, the singularity is a root for an irreducible factor of the leading coefficient of $C(z,w)$. In general, one may not be able to explicitly compute either the singularity ρ or the corresponding value a. In such cases, we rely on Puiseux approximations, which provide very accurate asymptotic estimates.

In Case I, after making the change of variable $s = 1 - \frac{z}{\rho}$, $w(s)$ has a Puiseux expansion at the singularity $s = 0$. In particular, $w(s)$ has the form $w(s) = a - g(s)s^{\alpha}$, for some $\alpha \in \mathbb{R}^+$, the first positive exponent of that expansion, and a function $g(s)$ such that $g(s) = b + h(s)s^{\beta}$, with $h(0) \neq 0$ and $b, \beta \in \mathbb{R}^+$. First, the value of α can be obtained with the help of the Newton's polygon technique. We refer the reader to Broda et al. [6] for a detailed description of this technique. For this paper, however, it holds that $\alpha = \frac{1}{2}$. For the value of b, we can use the fact that $g(0) = b$.

In Case II, by making $v(z) = \frac{1}{w(z)}$, one concludes as above that $v(s) = cs^{\alpha} - g(s)s^{\alpha+\beta}$, with $c, \beta \in \mathbb{R}^+$ and for some Puiseux series $g(s)$ with non-negative exponents. The value for c can be obtained in a similar fashion as for b in Case I.

With these values, we can apply the following result to obtain the asymptotic behaviour of the coefficients of $w(z)$:

Theorem 3 ([6]). *With the notations and in the conditions above, we have:*

$$[z^n]w(z) \underset{n \to \infty}{\sim} \begin{cases} \frac{-b}{\Gamma(-\alpha)}\rho^{-n}n^{-\alpha-1} & \text{if } \lim_{z\to p} w(z) = a; \\ \frac{1}{c\Gamma(\alpha)}\rho^{-n}n^{\alpha-1} & \text{if } \lim_{z\to p} w(z) = +\infty, \end{cases}$$

where a, b, c, α, and ρ can be computed as above described.

Here, we use the notation $\underset{n \to \infty}{\sim}$ to describe the asymptotic behaviour of a function as n grows to infinity. In the following sections, we shall denote the constants as a_k, b_k, c_k, and ρ_k, to stress their dependence on the size of the alphabet.

6.1 Finite Languages

Let Σ be a k-letter alphabet and let $A_k(z) = \sum_{\alpha} z^{\|\alpha\|}$, $E_k(z) = \sum_{\alpha_\varepsilon} z^{\|\alpha_\varepsilon\|}$ and $N_k(z) = \sum_{\alpha_{\overline{\varepsilon}}} z^{\|\alpha_{\overline{\varepsilon}}\|}$ be the generating functions for the number of α-expressions, α_ε-expressions, and $\alpha_{\overline{\varepsilon}}$-expressions, respectively. According to the specifications (2) and (3), we have

$$E_k(z) = zN_k(z) + zE_k(z)^2 + 2zE_k(z)N_k(z) \quad \text{and}$$
$$N_k(z) = kz + 2zE_k(z)N_k(z) + 2zN_k(z)^2.$$

Now, by (1), we have $A_k(z) = 2z + E_k(z) + N_k(z)$. By the techniques described above, we obtain an algebraic curve for $A_k(z)$. As the leading coefficient of the obtained curve does not have positive roots, we are in Case I. We then obtain asymptotic approximations for the values of ρ_k, a_k, and b_k, namely,

$$\rho_k \underset{k \to \infty}{\sim} \frac{2\sqrt{2k}-1}{8k}, \quad a_k \underset{k \to \infty}{\sim} \sqrt{\frac{k}{2}}, \quad b_k \underset{k \to \infty}{\sim} \sqrt{k}.$$

This gives us the following asymptotic behaviour for the coefficients of $A_k(z)$*:*

$$[z^n]A_k(z) \underset{n,\,k\,\to\,\infty}{\sim} \frac{2^{\frac{3n}{2}-1}k^{\frac{n+1}{2}}}{\sqrt{\pi}}n^{-\frac{3}{2}}. \tag{9}$$

Now, we shall analyse the size of $\pi(\alpha)$*, for* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$*. From Definition 2 and (7), upper bounds* $p'(\alpha_\varepsilon)$ *and* $p''(\alpha_{\overline{\varepsilon}})$ *for the number of elements in* $\partial^+(\alpha_\varepsilon)$ *and* $\partial^+(\alpha_{\overline{\varepsilon}})$*, respectively, satisfy:*

$$\begin{aligned} p'(\alpha_{\overline{\varepsilon}}^{?}) &= p''(\alpha_{\overline{\varepsilon}}), & p'(\alpha_\varepsilon + \alpha_{\overline{\varepsilon}}) &= p'(\alpha_{\overline{\varepsilon}} + \alpha_\varepsilon) = p'(\alpha_\varepsilon) + p''(\alpha_{\overline{\varepsilon}}) - 1, \\ p'(\alpha_\varepsilon \cdot \beta_\varepsilon) &= p'(\alpha_\varepsilon) + p'(\beta_\varepsilon), & p'(\alpha_\varepsilon + \beta_\varepsilon) &= p'(\alpha_\varepsilon) + p'(\beta_\varepsilon) - 1, \end{aligned}$$

and

$$\begin{aligned} p''(a) &= 1, & p''(\alpha_{\overline{\varepsilon}} + \beta_{\overline{\varepsilon}}) &= p''(\alpha_{\overline{\varepsilon}}) + p''(\beta_{\overline{\varepsilon}}) - 1, \\ p''(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) &= p''(\alpha_{\overline{\varepsilon}}) + p''(\beta_{\overline{\varepsilon}}), & p''(\alpha_{\overline{\varepsilon}} \cdot \alpha_\varepsilon) &= p''(\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) = p''(\alpha_{\overline{\varepsilon}}) + p'(\alpha_\varepsilon). \end{aligned}$$

These upper bounds correspond to the case where all unions are disjoint. Let $P'_k(z) = \sum_{\alpha_\varepsilon} p'(\alpha_\varepsilon)z^{\|\alpha_\varepsilon\|}$ *and* $P''_k(z) = \sum_{\alpha_{\overline{\varepsilon}}} p''(\alpha_{\overline{\varepsilon}})z^{\|\alpha_{\overline{\varepsilon}}\|}$ *be the generating functions for the size of* $\partial^+(\alpha_\varepsilon)$ *and* $\partial^+(\alpha_{\overline{\varepsilon}})$ *for* α_ε*-expressions and* $\alpha_{\overline{\varepsilon}}$*-expressions, respectively. It follows that*

$$\begin{aligned} P'_k(z) &= \sum_{\alpha_\varepsilon} p'(\alpha_\varepsilon)\, z^{\|\alpha_\varepsilon\|} \\ &= \sum_{\alpha_{\overline{\varepsilon}}^{?}} p'(\alpha_{\overline{\varepsilon}}^{?})z^{\|\alpha_{\overline{\varepsilon}}^{?}\|} + \sum_{\alpha_\varepsilon,\beta_\varepsilon} p'(\alpha_\varepsilon + \beta_\varepsilon)z^{\|\alpha_\varepsilon+\beta_\varepsilon\|} + \sum_{\alpha_\varepsilon,\alpha_{\overline{\varepsilon}}} p'(\alpha_\varepsilon + \alpha_{\overline{\varepsilon}})z^{\|\alpha_\varepsilon+\alpha_{\overline{\varepsilon}}\|} \\ &\quad + \sum_{\alpha_{\overline{\varepsilon}},\alpha_\varepsilon} p'(\alpha_{\overline{\varepsilon}} + \alpha_\varepsilon)z^{\|\alpha_{\overline{\varepsilon}}+\alpha_\varepsilon\|} + \sum_{\alpha_{\overline{\varepsilon}},\beta_\varepsilon} p'(\alpha_{\overline{\varepsilon}} \cdot \beta_\varepsilon)z^{\|\alpha_{\overline{\varepsilon}}\cdot\beta_\varepsilon\|} \\ &= zP''_k(z) + \left(2zP'_k(z)E_k(z) - zE_k(z)^2\right) \\ &\quad + 2\left(zP'_k(z)N_k(z) + zP''_k(z)E_k(z) - zE_k(z)N_k(z)\right) + 2zP'_k(z)E_k(z) \\ &= zP''_k(z) + 4zP'_k(z)E_k(z) + 2zP'_k(z)N_k(z) + 2zP''_k(z)E_k(z) \\ &\quad - 2zE_k(z)N_k(z) - zE_k(z)^2. \end{aligned}$$

and, analogously,

$$P''_k(z) = kz + 4zP''_k(z)N_k(z) + 2zP'_k(z)N_k(z) + 2zP''_k(z)E_k(z) - zN_k(z)^2.$$

Now, the generating function for the size of $\pi(\alpha)$*, with* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$*, namely* $P_k(z)$*, satisfies* $P_k(z) = P'_k(z) + P''_k(z)$*. In similar fashion to the case of* $A_k(z)$*, we can obtain an algebraic curve for* $P_k(z)$*. The curve is now in Case II, and has the same dominant singularity as* $A_k(z)$*. Then, as above described, we obtain an asymptotic approximation for* c_k*, yielding us* $c_k \underset{k\,\to\,\infty}{\sim} \frac{8}{\sqrt{k}}$*. This gives us the following asymptotic behaviour for the coefficients of* $P_k(z)$*:*

$$[z^n]P_k(z) \underset{n,\,k\,\to\,\infty}{\sim} \frac{2^{\frac{3n}{2}-3}k^{\frac{n+1}{2}}}{\sqrt{\pi}}n^{-\frac{1}{2}}. \tag{10}$$

Finally, let us obtain asymptotic estimates for the number of letters in an expression $\alpha \in \mathsf{RE}_{\text{fin}}$ *of size* n*, which is an upper bound for the number of states in the partial derivative automaton (Lemma 6). Let* $\ell'(\alpha_\varepsilon) = |\alpha_\varepsilon|_\Sigma$ *and* $\ell''(\alpha_{\overline{\varepsilon}}) = |\alpha_{\overline{\varepsilon}}|_\Sigma$ *be the number of letters in an expression* α_ε *and* $\alpha_{\overline{\varepsilon}}$*, respectively. One has* $\ell'(\varepsilon) = 0$, $\ell''(a) = 1$, $\ell'(\alpha_{\overline{\varepsilon}}^{?}) = \ell''(\alpha_{\overline{\varepsilon}})$, $\ell''(\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) = \ell'(\alpha_\varepsilon) + \ell''(\alpha_{\overline{\varepsilon}})$*, etc. Let,* $L'_k(z) = \sum_{\alpha_\varepsilon} \ell'(\alpha_\varepsilon) z^{\|\alpha_\varepsilon\|}$ *and* $L''_k(z) = \sum_{\alpha_{\overline{\varepsilon}}} \ell''(\alpha_{\overline{\varepsilon}}) z^{\|\alpha_{\overline{\varepsilon}}\|}$*. It follows that*

$$L'_k(z) = zL''_k(z) + 4zL'_k(z)E_k(z) + 2zL'_k(z)N_k(z) + 2zL''_k(z)E_k(z) \quad \text{and}$$
$$L''_k(z) = kz + 4zL''_k(z)N_k(z) + 2zL'_k(z)N_k(z) + 2zL''_k(z)E_k(z).$$

Also, let $L_k(z) = L'_k(z) + L''_k(z)$ *be the generating function for the number of letters in an expression* $\alpha \in \mathsf{RE}_{\text{fin}}$*. Again, the curve for* $L_k(z)$ *is in Case II with dominant singularity* ρ_k*, and, similarly to* $P_k(z)$*, we obtain an approximation for the constant* c'_k*, namely,* $c'_k \underset{k\to\infty}{\sim} \frac{4}{\sqrt{k}}$*. Thus, one obtains the following asymptotic estimate for the coefficients of* $L_k(z)$*:*

$$[z^n]L_k(z) \underset{n,\,k\to\infty}{\sim} \frac{2^{\frac{3n}{2}-2}k^{\frac{n+1}{2}}}{\sqrt{\pi}} n^{-\frac{1}{2}}. \tag{11}$$

Given the results in (9) and (11), the average number of letters in an expression $\alpha \in \mathsf{RE}_{\text{fin}}$ *of size* n*, denoted by* avL*, is asymptotically given by*

$$avL = \frac{[z^n]L_k(z)}{[z^n]A_k(z)} \underset{n,\,k\to\infty}{\sim} \frac{n}{2}.$$

With the result in (10), the average size of $\pi(\alpha)$ *for an expression* $\alpha \in \mathsf{RE}_{\text{fin}}$ *of size* n*, denoted by* avP*, is asymptotically given by*

$$avP = \frac{[z^n]P_k(z)}{[z^n]A_k(z)} \underset{n,\,k\to\infty}{\sim} \frac{n}{4}.$$

By plugging these values together we obtain:

Proposition 7. *For large values of* n *and* k*, an upper bound for the average number of states of* $\mathcal{A}_{\mathsf{PD}}$ *for expressions* $\alpha \in \mathsf{RE}_{\text{fin}}$ *of size* n *is* $\frac{|\alpha|_\Sigma}{2}$*.*

6.2 Shuffle Ideals

Now, we shall estimate the average number of states in the partial derivative automaton for expressions $\tau \in \mathsf{RE}_{\text{SI}}$*. Let* $T_k(z) = \sum_\tau z^{\|\tau\|}$, $B_k(z) = \sum_{\tau_{LR}} z^{\|\tau_{LR}\|}$*, and* $R_k(z) = \sum_{\tau_R} z^{\|\tau_R\|}$ *be the generating functions for the number of* τ*-expressions,* τ_{LR}*-expressions, and* τ_R*-expressions, respectively. From (5) and (6), respectively, we have*

$$B_k(z) = kz^3 + zB_k(z)^2 + zB_k(z)R_k(z) \quad \text{and} \quad R_k(z) = kz^2 + 2zR_k(z)^2.$$

Finally, from (4), we have $T_k(z) = 2z + B_k(z)$*. The algebraic curve for* $T_k(z)$ *is in Case I, so we apply the same techniques as in the case of* $A_k(z)$*, to obtain* *asymptotic approximations for the values of $\overline{\rho}_k$, $\overline{a}_k$, and $\overline{b}_k$. In this case, we obtain*

$$\overline{\rho}_k = \frac{1}{2k^{\frac{1}{3}}}, \quad \overline{a}_k \underset{k \to \infty}{\sim} \frac{4455k^{\frac{2}{3}} + 18k^{\frac{1}{3}} + 729k + 5}{4374k}, \quad \overline{b}_k \underset{k \to \infty}{\sim} \frac{1}{6\sqrt{3}}.$$

Then, we get the following asymptotic behaviour for the coefficients of $T_k(z)$:

$$[z^n]T_k(z) \underset{n,\, k \to \infty}{\sim} \frac{2^{n-2}k^{\frac{n}{3}}}{3\sqrt{3\pi}} n^{-\frac{3}{2}}. \tag{12}$$

As we did for the case of finite languages, we shall now estimate the size of $\pi(\tau)$ for $\tau \in \mathsf{RE}_{\mathrm{SI}}$. From Definition 3 and (8), upper bounds $q'(\tau_{LR})$ and $q''(\tau_R)$ for the number of elements in $\partial^+(\tau_{LR})$ and $\partial^+(\tau_R)$, respectively, satisfy:

$$\begin{aligned} q'(\Sigma^\star\, a\, \Sigma^\star) &= 2, & q''(a\, \Sigma^\star) &= 1, \\ q'(\tau_{LR} + \tau'_{LR}) &= q'(\tau_{LR}) + q'(\tau'_{LR}) - 1, & q''(\tau_R + \tau'_R) &= q''(\tau_R) + q''(\tau'_R) - 1, \\ q'(\tau_{LR} \cdot \tau_R) &= q'(\tau_{LR}) + q''(\tau_R), & q''(\tau_R \cdot \tau_R) &= q''(\tau_R) + q''(\tau_R). \end{aligned}$$

Let $Q'_k(z) = \sum_{\tau_{LR}} q'(\tau_{LR}) z^{\|\tau_{LR}\|}$ and $Q''_k(z) = \sum_{\tau_R} q''(\tau_R) z^{\|\tau_R\|}$ be the generating functions for the size of $\partial^+(\tau_{LR})$ and $\partial^+(\tau_R)$ for τ_{LR}-expressions and τ_R-expressions, respectively. It follows that

$$Q'_k(z) = 2kz^3 + 2zQ'_k(z)B_k(z) + zQ'_k(z)R_k(z) + zQ''_k(z)B_k(z) - zB_k(z)^2$$

and $Q''_k(z) = kz^2 + 4zQ''_k(z)R_k(z) - zR_k(z)^2$. Now, the generating function for the size of $\pi(\tau)$, with $\tau \in \mathsf{RE}_{\mathrm{SI}}$, namely $Q_k(z)$, satisfies $Q_k(z) = z + Q'_k(z)$. The algebraic curve for $Q_k(z)$ is in the Case II, and, by a similar procedure as in the case of $P_k(z)$, we shall compute the constant $\overline{c}_k$. Here, we obtain $\overline{c}_k \underset{k \to \infty}{\sim} \sqrt{\frac{256}{k^{\frac{2}{3}}} - \frac{1728}{k^{\frac{1}{3}}} + 2592}$. This gives us the following asymptotic behaviour for the coefficients of $Q_k(z)$:

$$[z^n]Q_k(z) \underset{n,\, k \to \infty}{\sim} \frac{2^{n-\frac{7}{2}}k^{\frac{n}{3}-1}(54k^{\frac{2}{3}} + 19k^{\frac{1}{3}} + 162k + 7)}{729\sqrt{\pi}} n^{-\frac{1}{2}}. \tag{13}$$

Finally, let us obtain asymptotic estimates for the number of $\Sigma^\star$-expressions $\tau \in \mathsf{RE}_{\mathrm{SI}}$ of size n, which, for this grammar, is an upper bound for the number of states in the partial derivative automaton (Lemma 8). Let $s'(\tau_{LR}) = |\tau_{LR}|_{\Sigma^\star}$ and $s''(\tau_R) = |\tau_R|_{\Sigma^\star}$ be the number of $\Sigma^\star$-expressions in an expression τ_{LR} and τ_R, respectively. One has $s'(\Sigma^\star\, a\, \Sigma^\star) = 2$, $s''(a\, \Sigma^\star) = 1$, $s'(\tau_{LR} \cdot \tau_R) = s'(\tau_{LR}) + s''(\tau_R)$, $s''(\tau_R \cdot \tau'_R) = s''(\tau_R) + s''(\tau'_R)$, etc. Let $S'_k(z) = \sum_{\tau_{LR}} s'(\tau_{LR}) z^{\|\tau_{LR}\|}$ and $S''_k(z) = \sum_{\tau_R} s''(\tau_R) z^{\|\tau_R\|}$. It follows that

$$S'_k(z) = 2kz^3 + 2zS'_k(z)B_k(z) + zS'_k(z)R_k(z) + zS''_k(z)B_k(z)$$

and $S''_k(z) = kz^2 + 4zS''_k(z)R_k(z)$. Let also $S_k(z) = z + S'_k(z) + S''_k(z)$ be the generating function for the number of $\Sigma^\star$-expressions in an expression $\tau \in \mathsf{RE}_{\mathrm{SI}}$.

The curve for $S_k(z)$ is in the Case II, so we compute the value of the constant $\overline{c}'_k$, yielding us $\overline{c}'_k \underset{k\to\infty}{\sim} 2\sqrt{2}\sqrt{\frac{8}{k^{\frac{2}{3}}} - \frac{54}{k^{\frac{1}{3}}} + 81}$. With these values, we obtain the following asymptotic behaviour for the coefficients of $S_k(z)$:

$$[z^n]S_k(z) \underset{n,\,k\to\infty}{\sim} \frac{2^{n-\frac{5}{2}}k^{\frac{n}{3}-1}(54k^{\frac{2}{3}}+19k^{\frac{1}{3}}+162k+7)}{729\sqrt{\pi}}n^{-\frac{1}{2}}. \tag{14}$$

Given the results in (12) and (14), the average number of $\Sigma^\star$-expressions in an expression $\tau \in \mathsf{RE}_{\mathrm{SI}}$ of size n, denoted by avS, is asymptotically given by

$$avS = \frac{[z^n]S_k(z)}{[z^n]T_k(z)} \underset{n,\,k\to\infty}{\sim} \sqrt{\frac{2}{3}}n.$$

With the results in (13), the average size of $\pi(\tau)$ for an expression $\tau \in \mathsf{RE}_{\mathrm{SI}}$ of size n, denoted by avQ, is asymptotically given by

$$avQ = \frac{[z^n]Q_k(z)}{[z^n]T_k(z)} \underset{n,\,k\to\infty}{\sim} \frac{n}{\sqrt{6}}.$$

With these values together we obtain:

Proposition 8. *For large values of n and k, an upper bound for the average number of states of $\mathcal{A}_{\mathsf{PD}}$ for expressions $\tau \in \mathsf{RE}_{\mathrm{SI}}$ of size n is $\frac{|\tau|_{\Sigma^\star}}{2}$.*

6.3 Comparison of the Results

Let us now compare the results obtained for the two classes of expressions. We will start by evaluating the growth of the size of an expression $\alpha \in \mathsf{RE}_{\mathrm{fin}}$ and its respective $\tau \in \mathsf{RE}_{\mathrm{SI}}$ such that $\mathsf{e}(\alpha) = \tau$ and thus $\mathcal{L}(\tau) = \mathcal{L}(\alpha) \sqcup \Sigma^\star$. By Lemma 5, we have that $\|\mathsf{e}(\alpha)\| \le 2\|\alpha\| + 1$.

First, let $h(\alpha_{\overline{\varepsilon}}) = \|\overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}})\|$ denote the size of an expression $\overrightarrow{\mathsf{e}}(\alpha_{\overline{\varepsilon}})$, for $\alpha_{\overline{\varepsilon}}$-expressions. We have that $h(a) = 2$, $h(\alpha_{\overline{\varepsilon}} + \beta_{\overline{\varepsilon}}) = h(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) = h(\alpha_{\overline{\varepsilon}}) + h(\beta_{\overline{\varepsilon}}) + 1$, and $h(\alpha_{\overline{\varepsilon}} \cdot \alpha_\varepsilon) = h(\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) = h(\alpha_{\overline{\varepsilon}})$. Let $H_k(z) = \sum_{\alpha_{\overline{\varepsilon}}} h(\alpha_{\overline{\varepsilon}}) z^{\|\alpha_{\overline{\varepsilon}}\|}$. We have

$$H_k(z) = 2kz + 2zH_k(z)E_k(z) + 4zH_k(z)N_k(z) + 2zN_k(z)^2.$$

Now, let $g'(\alpha_\varepsilon) = \|\mathsf{e}(\alpha_\varepsilon)\|$ and $g''(\alpha_{\overline{\varepsilon}}) = \|\mathsf{e}(\alpha_{\overline{\varepsilon}})\|$. On the one hand, we have $g'(\alpha_\varepsilon) = 1$, for every α_ε-expression. On the other, we have $g''(a) = 3$, $g''(\alpha_{\overline{\varepsilon}} + \beta_{\overline{\varepsilon}}) = g''(\alpha_{\overline{\varepsilon}}) + g''(\beta_{\overline{\varepsilon}}) + 1$, $g''(\alpha_{\overline{\varepsilon}} \cdot \alpha_\varepsilon) = g''(\alpha_\varepsilon \cdot \alpha_{\overline{\varepsilon}}) = g''(\alpha_{\overline{\varepsilon}})$, and $g''(\alpha_{\overline{\varepsilon}} \cdot \beta_{\overline{\varepsilon}}) = g''(\alpha_{\overline{\varepsilon}}) + h(\beta_{\overline{\varepsilon}}) + 1$. Let $G'_k(z) = \sum_{\alpha_\varepsilon} g'(\alpha_\varepsilon) z^{\|\alpha_\varepsilon\|}$ and $G''_k(z) = \sum_{\alpha_{\overline{\varepsilon}}} g''(\alpha_{\overline{\varepsilon}}) z^{\|\alpha_{\overline{\varepsilon}}\|}$. We have $G'_k(z) = E_k(z)$ and

$$G''_k(z) = 3kz + 3zG''_k(z)N_k(z) + zG''_k(z)E_k(z) + 2zH_k(z)N_k(z) + 2zN_k(z)^2.$$

Finally, let $G_k(z) = 2z + G'_k(z) + G''_k(z)$ be the generating function for $\|\mathsf{e}(\alpha)\|$, with $\alpha \in \mathsf{RE}_{\mathrm{fin}}$. The algebraic curve for $G_k(z)$ is in the Case II, with the same dominant singularity as $A_k(z)$. Once again, we shall compute the value of c''_k by *the means above. Here, we obtain* $c''_k \underset{n,\,k\,\to\,\infty}{\sim} \frac{4}{3\sqrt{k}}$*. Then, we obtain the following asymptotic behaviour for the coefficients of* $G_k(z)$*:*

$$[z^n]G_k(z) \underset{n,\,k\,\to\,\infty}{\sim} \frac{3 \cdot 2^{\frac{3n}{2}-2}k^{\frac{n+1}{2}}}{\sqrt{\pi}} n^{-\frac{1}{2}}. \tag{15}$$

Thus, using the results in (9) and (15), the average size of $\mathsf{e}(\alpha)$ *for an expression* $\alpha \in \mathsf{RE}_{\mathrm{fin}}$ *of size* n*, denoted by* avG*, is asymptotically given by*

$$avG = \frac{[z^n]G_k(z)}{[z^n]A_k(z)} \underset{n,\,k\,\to\,\infty}{\sim} \frac{3}{2}n.$$

Then, taking into account the results for avP *and* avQ*, we obtain:*

Proposition 9. *The size of* $\mathcal{A}_{\mathsf{PD}}$ *for* $\mathsf{e}(\alpha)$ *is on average 2.45 times larger than the size of* $\mathcal{A}_{\mathsf{PD}}$ *for* α*.*

References

1. Angluin, D., Aspnes, J., Kontorovich, A.: On the learnability of shuffle ideals. In: Bshouty, N.H., Stoltz, G., Vayatis, N., Zeugmann, T. (eds.) Proc. 23rd ALT 2012. LNCS, vol. 7568, pp. 111–123. Springer (2012). https://doi.org/10.1007/978-3-642-34106-9_12
2. Antimirov, V.: Partial derivatives of regular expressions and finite automaton constructions. Theoretical Computer Science **155**(2), 291–319 (1996), https://doi.org/10.1016/0304-3975(95)00182-4
3. Birget, J.C.: Partial orders on words, minimal elements of regular languages, and state complexity. Theoretical Computer Science **119**(2), 267–291 (1993). https://doi.org/10.1016/0304-3975(93)90160-U
4. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: On the average number of states of partial derivative automata. In: Gao, Y., Lu, H., Seki, S., Yu, S. (eds.) Proc. 14th DLT 2010. LNCS, vol. 6224, pp. 112–123. Springer (2010). https://doi.org/10.1007/978-3-642-14455-4_12
5. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: A Hitchhiker's Guide to descriptional complexity through analytic combinatorics. Theoretical Computer Science **528**, 85–100 (2014). https://doi.org/10.1016/J.TCS.2014.02.013
6. Broda, S., Machiavelo, A., Moreira, N., Reis, R.: Guest column: Analytic combinatorics and descriptional complexity of regular languages on average. SIGACT News **51**(1), 38–56 (Mar 2020), https://doi.org/10.1145/3388392.3388401
7. Flajolet, P., Sedgewick, R.: Analytic combinatorics. CUP (2009)
8. Gruber, H., Holzer, M., Kutrib, M.: The size of Higman-Haines sets. Theoretical Computer Science **387**(2), 167–176 (Nov 2007). https://doi.org/10.1016/j.tcs.2007.07.036
9. Gruber, H., Holzer, M., Kutrib, M.: More on the size of Higman-Haines sets: Effective construction. Fundamenta Informaticae **91**(1), 105–121 (2009). https://doi.org/10.3233/FI-2009-0035
10. Héam, P.C.: On shuffle ideals. RAIRO–Informatique théorique et Applications / Theoretical Informatics and Applications **36**(4), 359–384 (2002). https://doi.org/10.1051/ita:2003002

11. Higman, G.: Ordering by divisibility in abstract algebras. Proceedings of the London Mathematical Society **3**(1), 326–336 (1952)
12. Karandikar, P., Niewerth, M., Schnoebelen, P.: On the state complexity of closures and interiors of regular languages with subwords and superwords. Theoretical Computer Science **610**, 91–107 (2016). https://doi.org/10.1016/j.tcs.2015.09.028
13. Mirkin, B.G.: An algorithm for constructing a base in a language of regular expressions. Engineering Cybernetics **5**, 51—57 (1966)
14. Pin, J.É.: Varieties of Formal Languages. North Oxford (1986)
15. Rogers, J., Heinz, J., Bailey, G., Edlefsen, M., Visscher, M., Wellcome, D., Wibel, S.: On languages piecewise testable in the strict sense. In: Ebert, C., Jäger, G., , Michaelis, J. (eds.) 10th and 11th MOL (2010). LNCS, vol. 6149, pp. 255–265. Springer (2010). https://doi.org/10.1007/978-3-642-14322-9_19
16. Straubing, H.: Finite semigroup varieties of the form $\mathbf{V} * \mathbf{D}$. Journal of Pure and Applied Algebra **36**(1), 53–94 (1985). https://doi.org/10.1016/0022-4049(85)90062-3
17. Thérien, D.: Classification of finite monoids: the language approach. Theoretical Computer Science **14**(2), 195–208 (1981). https://doi.org/10.1016/0304-3975(81)90057-8