Submission to the Journal of Automata, Languages and Combinatorics
Submitted on January 2, 2018

ON THE AVERAGE COMPLEXITY OF PARTIAL DERIVATIVE AUTOMATA FOR SEMI-EXTENDED EXPRESSIONS

Rafaela Bastos Sabine Broda António Machiavelo Nelma Moreira
Rogério Reis

CMUP & DCC, Faculdade de Ciências da Universidade do Porto
Rua do Campo Alegre 1024, 4250-007
Porto, Portugal
{rrbastos,sbb}@dcc.fc.up.pt ajmachia@fc.up.pt {nam,rvr}@dcc.fc.up.pt

ABSTRACT

Extended regular expressions (with complement and intersection) are used in many applications due to their succinctness. In particular, regular expressions extended with intersection only (also called semi-extended) can already be exponentially smaller than standard regular expressions or equivalent nondeterministic finite automata. For practical purposes it is important to study the average behaviour of conversions between these models. In this paper, we focus on the conversion of regular expressions with intersection to nondeterministic finite automata, using partial derivatives and the notion of support. We give a tight upper bound of $2^{O(n)}$ for the worst-case number of states of the resulting partial derivative automaton, where n is the size of the expression. Using the framework of analytic combinatorics, we establish an upper bound of $(1.056+o(1))^n$ for its asymptotic average-state complexity, which is significantly smaller than the one for the worst case. Some experimental results here presented suggest that, on average, the upper bound may not be exponential. Finally, we study the class of semi-extended regular expressions with only one occurrence of intersection at the top level. In this case, the worst-case state complexity of the partial derivative automaton is quadratic on the size of the expression, but we obtained an upper-bound that is, asymptotically and on average, $O(n^{\frac{3}{2}})$.

Keywords: regular expressions, intersection, automata, partial derivatives, average case complexity, analytic combinatorics

This work was partially supported by CMUP (UID/MAT/00144/2013), which is funded by FCT (Portugal) with national (MEC) and european structural funds through the programs FEDER, under the partnership agreement PT2020.

1. Introduction

Regular expressions with additional operators are used in applications such as programming languages [15], XML processing [29], or runtime verification [28]. Most of these operators do not increase their language expressive power but lead to gains in the succinctness of the representation. This is the case for intersection. For regular expressions with intersection ($\mathsf{RE}_\cap$) (or semi-extended), several computational complexity decision problems, such as membership, equivalence and emptiness, were studied by various authors. Petersen [27] has shown that the membership problem is LOGCFL-complete, while for standard regular expressions (RE) it is NL-complete [23]. Fürer [18] has proved that inequivalence and non-empty complement are EXPSPACE-complete, which contrasts with the PSPACE-completeness of these problems for RE. The complexity of the conversions from regular expressions with intersection to standard regular expressions, and to finite automata, were recently studied by Gelade and Neven [20], Gruber and Holzer [22], and Gelade [19]. The conversion from $\mathsf{RE}_\cap$ to RE, or to nondeterministic finite automata (NFA), is exponential and, it is double exponential to deterministic finite automata (DFA). The conversion from $\alpha \in \mathsf{RE}_\cap$ to a DFA can be accomplished using Brzozowski's derivatives [11]. From RE to NFA a standard conversion algorithm is the partial derivative automaton construction ($\mathcal{A}_{pd}$) introduced by Antimirov [1], which coincides with the resolution of systems of equations by Mirkin [24]. The average complexity of these conversions was recently studied using the framework of analytic combinatorics [6, 7], and also their extension to regular expressions with shuffle [9]. For these studies Mirkin's construction is essential, as it provides inductive definitions that can be used to obtain generating functions.

Caron et al. [12] extended the $\mathcal{A}_{pd}$ to regular expressions with both intersection and complement (extended regular expressions)[1]. In their approach, a partial derivative is a set of sets of expressions (akin a disjunctive normal form), whereas in the present work it is simply a set of expressions. In the worst-case, their approach also leads to NFAs that can be exponentially larger than the original expressions. Moreover, considering sets of sets of expressions would turn the analytic combinatoric analysis much harder.

In this paper we show that for $\mathsf{RE}_\cap$, Mirkin's construction can lead to automata not initially connected, and thus larger than the ones built by Antimirov's construction, although the two constructions can, in some cases, produce identical NFAs. Here we present an exponential worst-case upper bound which is tight for both. Using the framework of analytic combinatorics, we give an upper bound for the asymptotic average-state complexity for Mirkin's construction, which turns out to be much smaller than the worst-case one. This also means that Antimirov's construction is asymptotically, and on average, much smaller than the worst-case upper bound. Finally, we study a restricted family of semi-extended regular expressions with only one occurrence of intersection at the top level. In this case, the worst-case state complexity of the partial derivative automaton is quadratic in the size of the expression,

[1] And a more general framework is also reported in [13]. Similar approaches were considered by Bastos [3].

but asymptotically, and on average, it is $O(n^{\frac{3}{2}})$. A preliminary version of this paper appeared in [4].

The rest of the paper is organised as follows. In Section 2 we introduce the set of regular expressions with intersection and recall its algebraic structure. In Section 3 we define the notion of support as a solution of a system of linear expression equations, and study its size in the worst case. In Section 4 we consider partial derivatives for semi-extended expressions and the partial derivative automaton, $\mathcal{A}_{pd}$, and we also show that the set of partial derivatives of an expression α (w.r.t. non-empty words) can be a proper subset of the support of α. In Section 5 we use the framework of analytic combinatorics to obtain several average results. First, we present parametrised generating functions for regular expressions with any number of unary and binary operators over a given alphabet. These are used to obtain the asymptotic average number of intersections, as well as the number of letters in semi-extended expressions of a given size. Then, we calculate an upper bound for the asymptotic average size of the support, which also provides an upper bound for the average state complexity of $\mathcal{A}_{pd}$. We present some experimental results suggesting that, on average, the size is much smaller. In Section 6 we study regular expressions with only one occurrence of intersection at the top level. We present our conclusions in Section 7.

2. Regular Expressions with Intersection

Let $\Sigma = \{a_1, \dots, a_k\}$ be an *alphabet* of size k. A *word* over Σ is a finite sequence of symbols of Σ. The *empty word* is denoted by ε. The set $\Sigma^\star$ is the set of all words over Σ, and Σ^+ denotes $\Sigma^\star \setminus \{\varepsilon\}$. A *language* over Σ is a subset of $\Sigma^\star$. The set $\mathsf{RE}_\cap$ of *regular expressions with intersection* over Σ contains the expression $\emptyset$, and all terms generated by the following grammar:

$$\alpha \rightarrow \varepsilon \mid a \mid (\alpha + \alpha) \mid (\alpha \cdot \alpha) \mid (\alpha \cap \alpha) \mid (\alpha^\star) \qquad (a \in \Sigma), \tag{1}$$

where the operator $\cdot$ (*concatenation*) is often omitted. Parentheses can also be omitted considering the following precedences for the operators: $\star > \cdot > \cap > +$. We denote by RE the set of standard expressions, where $\cap$ does not occur. The size of a regular expression $\alpha \in \mathsf{RE}_\cap$ is denoted by $|\alpha|$, and defined as the number of occurrences of symbols (parenthesis not counted) in α. Similarly, $|\alpha|_\Sigma$ denotes the number of occurrences of alphabet symbols (letters) in α, and $|\alpha|_\cap$ the number of occurrences of the binary operator $\cap$. The language $\mathcal{L}(\alpha)$ for $\alpha \in \mathsf{RE}_\cap$ is defined as usual, adding $\mathcal{L}(\alpha \cap \beta) = \mathcal{L}(\alpha) \cap \mathcal{L}(\beta)$. We say that two regular expressions $\alpha, \beta \in \mathsf{RE}_\cap$ are *equivalent*, if $\mathcal{L}(\alpha) = \mathcal{L}(\beta)$, and write $\alpha \doteq \beta$ in this case. For a set $S \subseteq \mathsf{RE}_\cap$, the language of S is defined as $\mathcal{L}(S) = \bigcup_{\alpha \in S} \mathcal{L}(\alpha)$. The notion of equivalence extends naturally to sets of regular expressions. The *left-quotient* of a language $\mathcal{L}$ w.r.t. a word $w \in \Sigma^\star$ is defined as $w^{-1}\mathcal{L} = \{\ x \mid wx \in \mathcal{L}\ \}$. The algebraic structure $(\mathsf{RE}_\cap, +, \cdot, \emptyset, \varepsilon)$ constitutes an idempotent semiring, which with the unary operator $\star$ is a Kleene algebra. Antimirov and Mosses [2] presented a complete and sound axiomatisation for $\mathsf{RE}_\cap$, where the binary operator $\cap$ is idempotent, commutative, associative, distributes over $+$, and also satisfies the following axioms, where $a_i, a_j \in \Sigma$ and $\alpha, \beta, \gamma \in \mathsf{RE}_\cap$:

$$\begin{array}{ll}
(\varepsilon \cap \beta) \doteq \emptyset \wedge (\alpha \doteq \beta\alpha + \gamma) \Rightarrow \alpha \doteq \beta^\star\gamma, & \varepsilon \cap \alpha^\star \doteq \varepsilon, \\
\varepsilon \cap (\alpha\beta) \doteq (\varepsilon \cap \alpha) \cap \beta, & \varepsilon \cap a_i \doteq \emptyset \cap \alpha \doteq \emptyset, \\
(a_i\alpha) \cap (a_j\beta) \doteq (a_i \cap a_j)(\alpha \cap \beta), & a_i \cap a_j \doteq \emptyset \quad (a_i \neq a_j), \\
(\alpha a_i) \cap (\beta a_j) \doteq (\alpha \cap \beta)(a_i \cap a_j), & \alpha + (\alpha \cap \beta) \doteq \alpha.
\end{array}$$

With the usual abuse of notation, define the function $\varepsilon : \mathsf{RE}_\cap \to \{\emptyset, \varepsilon\}$ by $\varepsilon(\alpha) = \varepsilon$ if $\varepsilon \in \mathcal{L}(\alpha)$, and $\varepsilon(\alpha) = \emptyset$ otherwise. The methods developed in Sections 3 and 4 are syntactical, and aim at building automata equivalent to a given regular expression. To ensure the finiteness of the constructions it is not necessary to consider regular expressions modulo any of the above properties[2]. However, in some examples, for the sake of succinctness, we also consider regular expressions modulo the identities of $\cdot$ and $+$. Note that this does not affect the upper bounds of the number of states, both in the worst and in the average case.

3. Automata and Systems of Equations

We first recall the definition of a nondeterministic finite automaton (NFA) as a tuple $\mathcal{A} = \langle S, \Sigma, S_0, \delta, F\rangle$, where S is a finite set of states, Σ is a finite alphabet, $S_0 \subseteq S$ a set of initial states, $\delta : S \times \Sigma \to 2^S$ the transition function, and $F \subseteq S$ a set of final states. This function δ can be naturally extended to sets of states and to words. In what follows we will take $S = [1, n]$. The *language* of $\mathcal{A}$ is $\mathcal{L}(\mathcal{A}) = \{\ w \in \Sigma^\star \mid \delta(S_0, w) \cap F \neq \emptyset\ \}$. The *right language* of a state s, denoted by $\mathcal{L}_s$, is the language accepted by $\mathcal{A}$ if we take $S_0 = \{s\}$. It is well known that it is possible to associate to each n-state NFA $\mathcal{A}$ over $\Sigma = \{a_1, \dots, a_k\}$, with right languages $\mathcal{L}_1, \dots, \mathcal{L}_n$, a system of linear language equations

$$\mathcal{L}_i = a_1\mathcal{L}_{1i} \cup \cdots \cup a_k\mathcal{L}_{ki} \cup \varepsilon(\mathcal{L}_i), \text{ for } i \in S,$$

where $\mathcal{L}_{ji} = \bigcup_{l \in \delta(i, a_j)} \mathcal{L}_l$ and $\mathcal{L}(\mathcal{A}) = \bigcup_{i \in S_0} \mathcal{L}_i$. In the same way, it is possible to associate to each regular expression a system of equations. We here extend Mirkin's contruction to regular expressions with intersection.

Definition 1. Consider $\alpha_0 \in \mathsf{RE}_\cap$ over $\Sigma = \{a_1, \dots, a_k\}$. A *support* of α_0 is a set $\{\alpha_1, \dots, \alpha_n\}$ of regular expressions with intersection that satisfies a system of equations

$$\alpha_i \doteq a_1\alpha_{1i} + \cdots + a_k\alpha_{ki} + \varepsilon(\alpha_i) \qquad i \in [0, n], \tag{2}$$

for some $\alpha_{1i}, \dots, \alpha_{ki}$, where each α_{ji} is a (possibly empty) sum of elements in $\{\alpha_1, \dots, \alpha_n\}$.

[2] As is the case, for instance, for Brzozowski's DFA or the approach of Caron et al.

It is clear that the existence of a support of α implies the existence of an NFA that accepts the language of α. A support for a regular expression $\alpha \in \mathsf{RE}_\cap$ can be computed using the function $\pi : \mathsf{RE}_\cap \to 2^{\mathsf{RE}_\cap}$ defined below. First, we define some operations on sets of regular expressions. Given $S, T \subseteq \mathsf{RE}_\cap$ and $\beta \in \mathsf{RE}_\cap$, we set $S\beta = \{\ \alpha\beta \mid \alpha \in S\ \}$ and $S \mathbin{\dot{\cap}} T = \{\ \alpha \cap \beta \mid \alpha \in S, \beta \in T\ \}$. Note, in particular, that $\mathcal{L}(S \mathbin{\dot{\cap}} T) = \mathcal{L}(S) \cap \mathcal{L}(T)$.

Definition 2. Given $\alpha \in \mathsf{RE}_\cap$, the set $\pi(\alpha)$ is inductively defined by:

$$\begin{aligned} \pi(\emptyset) = \pi(\varepsilon) &= \emptyset, & \pi(\alpha+\beta) &= \pi(\alpha) \cup \pi(\beta),\\ \pi(a) &= \{\varepsilon\} \quad (a \in \Sigma), & \pi(\alpha\beta) &= \pi(\alpha)\beta \cup \pi(\beta),\\ \pi(\alpha^\star) &= \pi(\alpha)\alpha^\star, & \pi(\alpha \cap \beta) &= \pi(\alpha) \mathbin{\dot{\cap}} \pi(\beta). \end{aligned}$$

Example 3. Given the regular expression $\alpha_1 = (b + ab + aab + abab) \cap (ab)^\star$,

$$\begin{aligned} \pi(\alpha_1) = \{&bab \cap b(ab)^\star,\ ab \cap b(ab)^\star,\ b \cap b(ab)^\star,\ \varepsilon \cap b(ab)^\star,\ bab \cap (ab)^\star,\\ &ab \cap (ab)^\star,\ b \cap (ab)^\star,\ \varepsilon \cap (ab)^\star\}. \end{aligned}$$

Proposition 4. *If $\alpha \in \mathsf{RE}_\cap$, then $\pi(\alpha)$ is a support of α.*

Proof. We will proceed by induction on the structure of α. The proof for all cases, excluding $\alpha \cap \beta$, can be found in [24, 14, 6]. Let $\pi(\alpha_0) = \{\alpha_1, \dots, \alpha_n\}$ and $\pi(\beta_0) = \{\beta_1, \dots, \beta_m\}$ be a support of α_0 and β_0, respectively. Thus,

$$\alpha_i \doteq a_1\alpha_{1i} + \cdots + a_k\alpha_{ki} + \varepsilon(\alpha_i), \quad \text{for } i = 0, \dots, n$$

and

$$\beta_j \doteq a_1\beta_{1j} + \cdots + a_k\beta_{kj} + \varepsilon(\beta_j), \ \text{for } j = 0, \dots, m,$$

where, for all $l = 1, \dots, k$, α_{li} and β_{lj} are linear combinations of elements of $\pi(\alpha_0)$ and $\pi(\beta_0)$, respectively. We want to prove that $\pi(\alpha_0 \cap \beta_0)$ is a support for $\alpha_0 \cap \beta_0$. For $i = 0, \dots, n$ and $j = 0, \dots, m$, and using the axioms for $\cap$, we have

$$\begin{aligned} \alpha_i \cap \beta_j &\doteq (a_1\alpha_{1i} + \cdots + a_k\alpha_{ki} + \varepsilon(\alpha_i)) \cap (a_1\beta_{1j} + \cdots + a_k\beta_{kj} + \varepsilon(\beta_j))\\ &\doteq (a_1\alpha_{1i} \cap a_1\beta_{1j}) + \cdots + (a_1\alpha_{1i} \cap a_k\beta_{kj}) + (a_1\alpha_{1i} \cap \varepsilon(\beta_j)) + \cdots\\ &\quad \cdots + (a_k\alpha_{ki} \cap a_1\beta_{1j}) + \cdots + (a_k\alpha_{ki} \cap a_k\beta_{kj}) + (a_k\alpha_{ki} \cap \varepsilon(\beta_j)) + \cdots\\ &\quad \cdots + (\varepsilon(\alpha_i) \cap a_1\beta_{1j}) + \cdots + (\varepsilon(\alpha_i) \cap a_k\beta_{kj}) + (\varepsilon(\alpha_i) \cap \varepsilon(\beta_j))\\ &\doteq (a_1 \cap a_1)(\alpha_{1i} \cap \beta_{1j}) + \cdots + (a_k \cap a_k)(\alpha_{ki} \cap \beta_{kj}) + (\varepsilon(\alpha_i) \cap \varepsilon(\beta_j))\\ &\doteq a_1(\alpha_{1i} \cap \beta_{1j}) + \cdots + a_k(\alpha_{ki} \cap \beta_{kj}) + \varepsilon(\alpha_i \cap \beta_j). \end{aligned}$$

For each $l = 1, \dots, k$, we know that $\alpha_{li} = \sum_{i' \in I_{li}} \alpha_{i'}$ and $\beta_{lj} = \sum_{j' \in J_{lj}} \beta_{j'}$, for $I_{li} \subseteq \{1, \dots, n\}$ and $J_{lj} \subseteq \{1, \dots, m\}$. And, since

$$\alpha_{li} \cap \beta_{lj} \doteq \sum_{i' \in I_{li}} \alpha_{i'} \cap \sum_{j' \in J_{lj}} \beta_{j'} \doteq \sum_{i' \in I_{li}, j' \in J_{lj}} (\alpha_{i'} \cap \beta_{j'}),$$

we conclude that $\pi(\alpha_0) \mathbin{\dot{\cap}} \pi(\beta_0) = \{\alpha_1 \cap \beta_1, \dots, \alpha_1 \cap \beta_m, \dots, \alpha_n \cap \beta_m\}$ is a support for $\alpha_0 \cap \beta_0$. □

The next proposition provides an upper bound on the cardinality of the support of a regular expression.

Proposition 5. *For all* $\alpha \in \mathit{RE}_\cap$*, the inequality* $|\pi(\alpha)| \leq 2^{|\alpha|_\Sigma - |\alpha|_\cap - 1}$ *holds.*

Proof. We proceed by induction on the structure of the regular expression α. It is easily proved that the statement holds for the base cases ε, $\emptyset$ and $a \in \Sigma$. Assume that the result holds for some $\alpha, \beta \in \mathsf{RE}_\cap$. We will make use of the fact that $2^m + 2^n \leq 2^{m+n+1}$, for any $m, n \geq 0$. For $\alpha + \beta$, one has

$$\begin{aligned}
|\pi(\alpha+\beta)| &= |\pi(\alpha) \cup \pi(\beta)| \leq |\pi(\alpha)| + |\pi(\beta)| \\
&\leq 2^{|\alpha|_\Sigma - |\alpha|_\cap - 1} + 2^{|\beta|_\Sigma - |\beta|_\cap - 1} \\
&\leq 2^{|\alpha|_\Sigma - |\alpha|_\cap - 1 + |\beta|_\Sigma - |\beta|_\cap - 1 + 1} = 2^{|\alpha+\beta|_\Sigma - |\alpha+\beta|_\cap - 1}.
\end{aligned}$$

The case for $\alpha\beta$ is analogous. For $\alpha^\star$, one has

$$|\pi(\alpha^\star)| = |\pi(\alpha)\alpha^\star| = |\pi(\alpha)| \leq 2^{|\alpha|_\Sigma - |\alpha|_\cap - 1} = 2^{|\alpha^\star|_\Sigma - |\alpha^\star|_\cap - 1}.$$

Finally, for $\alpha \cap \beta$, one has

$$\begin{aligned}
|\pi(\alpha \cap \beta)| &= |\pi(\alpha) \mathbin{\dot{\cap}} \pi(\beta)| \\
&\leq |\pi(\alpha)| \cdot |\pi(\beta)| \leq 2^{|\alpha|_\Sigma - |\alpha|_\cap - 1} \cdot 2^{|\beta|_\Sigma - |\beta|_\cap - 1} \\
&= 2^{|\alpha \cap \beta|_\Sigma - |\alpha \cap \beta|_\cap - 1}.
\end{aligned}$$

□

The next examples present families of regular expressions that witness the tightness of the upper bound established in Proposition 5.

Example 6. Let the regular expression $r_n \in \mathsf{RE}_\cap$ over $\{a, b\}$ be inductively defined by $r_0 = a^\star b^\star$, $r_1 = b^\star a$ and $r_n = r_{n-2} \cap r^\star_{n-1}$, for $n \geq 2$. Using the definition of support it is straightforward that $|\pi(r_0)| = |\{a^\star b^\star, b^\star\}| = 2^1$, $|\pi(r_1)| = |\{b^\star a, \varepsilon\}| = 2^1$, and $|\pi(r_n)| = |\pi(r_{n-2})| \cdot |\pi(r_{n-1})|$, for $n \geq 2$. Thus, we obtain $|\pi(r_n)| = 2^{\mathsf{fib}(n)}$, for $n \geq 0$, where $\mathsf{fib}(n)$ is the nth term of the Fibonacci sequence. Also, $|r_0|_\Sigma - |r_0|_\cap - 1 = 1$, $|r_1|_\Sigma - |r_1|_\cap - 1 = 1$, and

$$\begin{aligned}
|r_n|_\Sigma - |r_n|_\cap - 1 &= |r_{n-2}|_\Sigma + |r_{n-1}|_\Sigma - |r_{n-2}|_\cap - |r_{n-1}|_\Sigma - 2 \\
&= (|r_{n-2}|_\Sigma - |r_{n-2}|_\cap - 1) + (|r_{n-1}|_\Sigma - |r_{n-1}|_\cap - 1),
\end{aligned}$$

for $n \geq 2$. Consequently, $|r_n|_\Sigma - |r_n|_\cap - 1 = \mathsf{fib}(n)$, for $n \geq 0$. We conclude that $|\pi(r_n)| = 2^{|r_n|_\Sigma - |r_n|_\cap - 1}$, for $n \geq 0$.

Example 7. Let the regular expression $r_n \in \mathsf{RE}_\cap$ over $\{a\}$, be defined inductively by $r_0 = a^\star a$ and $r_n = r_{n-1} \cap a^\star a$, for $n \geq 1$. We have $\pi(r_0) = \pi(a^\star a) = \{a^\star a, \varepsilon\}$, and for $n \geq 1$,

$$\pi(r_n) = \underbrace{\{a^* a, \varepsilon\} \mathbin{\dot{\cap}} \cdots \mathbin{\dot{\cap}} \{a^* a, \varepsilon\}}_{n+1}.$$

Thus $|\pi(r_0)| = 2$ and $|\pi(r_n)| = |\pi(r_0)|^{n+1} = 2^{n+1}$. Note that $|r_n|_\Sigma = 2n + 2$ and $|r_n|_\cap = n$. Therefore $|\pi(r_n)| = 2^{|r_n|_\Sigma - |r_n|_\cap - 1}$.

4. Partial Derivatives

The notions of partial derivatives and partial derivative automata were introduced by Antimirov [1] for standard regular expressions. We now consider the Antimirov construction from $\mathsf{RE}_\cap$ expressions to NFAs.

Definition 8. For a regular expression $\alpha \in \mathsf{RE}_\cap$ and a symbol $a \in \Sigma$, the set $\partial_a(\alpha)$ of *partial derivatives* of α w.r.t. a is defined by:

$$\begin{aligned} \partial_a(\emptyset) &= \emptyset, \\ \partial_a(\varepsilon) &= \emptyset, \\ \partial_a(b) &= \begin{cases} \{\varepsilon\}, & \text{if } a = b \\ \emptyset & \text{otherwise,} \end{cases} \end{aligned} \qquad \begin{aligned} \partial_a(\alpha\beta) &= \begin{cases} \partial_a(\alpha)\beta \cup \partial_a(\beta), & \text{if } \varepsilon(\alpha) = \varepsilon \\ \partial_a(\alpha)\beta & \text{otherwise,} \end{cases} \\ \partial_a(\alpha + \beta) &= \partial_a(\alpha) \cup \partial_a(\beta), \\ \partial_a(\alpha \cap \beta) &= \partial_a(\alpha) \mathbin{\dot\cap} \partial_a(\beta), \\ \partial_a(\alpha^\star) &= \partial_a(\alpha)\alpha^\star. \end{aligned}$$

This definition is extended to words $w \in \Sigma^\star$ by $\partial_\varepsilon(\alpha) = \{\alpha\}$, $\partial_{wa}(\alpha) = \bigcup_{\alpha_i \in \partial_w(\alpha)} \partial_a(\alpha_i)$, and $\partial_w(R) = \bigcup_{\alpha_i \in R} \partial_w(\alpha_i)$, where $R \subseteq \mathsf{RE}_\cap$. It easily follows that $\mathcal{L}(\partial_w(\alpha)) = w^{-1}\mathcal{L}(\alpha)$. The set of partial derivatives of an expression α is $\partial(\alpha) = \bigcup_{w \in \Sigma^\star} \partial_w(\alpha)$. We also define $\partial^+(\alpha) = \bigcup_{w \in \Sigma^+} \partial_w(\alpha)$.

As for standard regular expressions, the partial derivative automaton of an expression $\alpha \in \mathsf{RE}_\cap$ is defined by

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

where $F_\alpha = \{\ \gamma \in \partial(\alpha) \mid \varepsilon(\gamma) = \varepsilon\ \}$ and $\delta_\alpha(\gamma, a) = \partial_a(\gamma)$. It follows that $\mathcal{L}(\mathcal{A}_{pd}(\alpha))$ is exactly $\mathcal{L}(\alpha)$. Mirkin's and Antimirov's constructions coincide for standard regular expressions. We will see that this is not true for regular expressions with intersection.

The following lemmas present some properties of the function ∂_w, used to prove Proposition 11.

Lemma 9. *For all $S, S' \subseteq \mathit{RE}_\cap$ and $a \in \Sigma$, the following property holds*

$$\partial_a(S \mathbin{\dot\cap} S') = \partial_a(S) \mathbin{\dot\cap} \partial_a(S').$$

Proof. Let $a \in \Sigma$ and let $S, S' \subseteq \mathsf{RE}_\cap$, with $S = \{\alpha_1, \ldots, \alpha_n\}$ and $S' = \{\beta_1, \ldots, \beta_m\}$.

Then,

$$\begin{aligned}
\partial_a(S \mathbin{\dot\cap} S') &= \partial_a(\{\alpha_1,\dots,\alpha_n\} \mathbin{\dot\cap} \{\beta_1,\dots,\beta_m\}) \\
&= \partial_a(\{\alpha_1 \cap \beta_1,\dots,\alpha_1 \cap \beta_m,\dots,\alpha_n \cap \beta_1,\dots,\alpha_n \cap \beta_m\}) \\
&= \partial_a(\alpha_1 \cap \beta_1) \cup \dots \cup \partial_a(\alpha_1 \cap \beta_m) \cup \cdots \\
&\quad \cdots \cup \partial_a(\alpha_n \cap \beta_1) \cup \dots \cup \partial_a(\alpha_n \cap \beta_m) \\
&= (\partial_a(\alpha_1) \mathbin{\dot\cap} \partial_a(\beta_1)) \cup \dots \cup (\partial_a(\alpha_1) \mathbin{\dot\cap} \partial_a(\beta_m)) \cup \cdots \\
&\quad \cdots \cup (\partial_a(\alpha_n) \mathbin{\dot\cap} \partial_a(\beta_1)) \cup \dots \cup (\partial_a(\alpha_n) \mathbin{\dot\cap} \partial_a(\beta_m)) \\
&= \bigcup_{\alpha_i \in S, \beta_j \in S'} \{\alpha'_i \cap \beta'_j \mid \alpha'_i \in \partial_a(\alpha_i), \beta'_j \in \partial_a(\beta_j)\} \\
&= \bigcup_{\alpha_i \in S} \partial_a(\alpha_i) \mathbin{\dot\cap} \bigcup_{\beta_j \in S'} \partial_a(\beta_j) \\
&= \partial_a(S) \mathbin{\dot\cap} \partial_a(S').
\end{aligned}$$

□

Let $\mathrm{suff}(w)$ be the set of all non-empty suffixes of w, being defined as $\mathrm{suff}(w) = \{\, v \in \Sigma^+ \mid \exists u \in \Sigma^\star : uv = w \,\}$. Except for the second case, the following lemma was shown by Antimirov.

Lemma 10. *For every regular expressions $\alpha, \beta \in \mathsf{RE}_\cap$ and word $w \in \Sigma^+$, ∂_w satisfies the following:*

$$\partial_w(\alpha + \beta) = \partial_w(\alpha) \cup \partial_w(\beta), \tag{3}$$

$$\partial_w(\alpha \cap \beta) = \partial_w(\alpha) \mathbin{\dot\cap} \partial_w(\beta), \tag{4}$$

$$\partial_w(\alpha\beta) \subseteq \partial_w(\alpha)\beta \cup \bigcup_{v \in \mathrm{suff}(w)} \partial_v(\beta), \tag{5}$$

$$\partial_w(\alpha^\star) \subseteq \bigcup_{v \in \mathrm{suff}(w)} \partial_v(\alpha)\alpha^\star. \tag{6}$$

Proof. Antimirov[1] proved equations (3), (5) and (6). Thus, we only present the proof for equation (4).

The proof of the statement $\partial_w(\alpha \cap \beta) = \partial_w(\alpha) \mathbin{\dot\cap} \partial_w(\beta)$ is done by induction on w. If $w = \varepsilon$, then $\partial_\varepsilon(\alpha \cap \beta) = \{\alpha \cap \beta\} = \{\alpha\} \mathbin{\dot\cap} \{\beta\} = \partial_\varepsilon(\alpha) \mathbin{\dot\cap} \partial_\varepsilon(\beta)$. Suppose that $\partial_w(\alpha \cap \beta) = \partial_w(\alpha) \mathbin{\dot\cap} \partial_w(\beta)$ holds for a given w, then for wa, with $a \in \Sigma$, it follows from Lemma 9 that

$$\begin{aligned}
\partial_{wa}(\alpha \cap \beta) &= \partial_a(\partial_w(\alpha \cap \beta)) = \partial_a(\partial_w(\alpha) \mathbin{\dot\cap} \partial_w(\beta)) \\
&= \partial_a(\partial_w(\alpha)) \mathbin{\dot\cap} \partial_a(\partial_w(\beta)) = \partial_{wa}(\alpha) \mathbin{\dot\cap} \partial_{wa}(\beta).
\end{aligned}$$

□

Proposition 11. *For every regular expressions $\alpha, \beta \in \mathsf{RE}_\cap$, the following holds.*

$$\begin{aligned}
\partial^+(\alpha + \beta) &\subseteq \partial^+(\alpha) \cup \partial^+(\beta), & \partial^+(\alpha \cap \beta) &\subseteq \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta), \\
\partial^+(\alpha\beta) &\subseteq \partial^+(\alpha)\beta \cup \partial^+(\beta), & \partial^+(\alpha^\star) &\subseteq \partial^+(\alpha)\alpha^\star.
\end{aligned}$$

Proof. First note that, given a set $E \subseteq \mathsf{RE}_\cap$ and a regular expression $\alpha \in \mathsf{RE}_\cap$, if, for all $w \in \Sigma^+$, we have that $\partial_w(\alpha) \subseteq E$, then we have $\bigcup_{w\in\Sigma^+} \partial_w(\alpha) \subseteq E$ and thus $\partial^+(\alpha) \subseteq E$. Moreover, we know that for every $w \in \Sigma^+$, $\partial_w(\alpha) \subseteq \partial^+(\alpha)$. Let $\alpha, \beta \in \mathsf{RE}_\cap$ be regular expressions over Σ. Now,

- From equation (3), for all $w \in \Sigma^+$, the following holds:
$$\partial_w(\alpha + \beta) = \partial_w(\alpha) \cup \partial_w(\beta) \subseteq \partial^+(\alpha) \cup \partial^+(\beta),$$
and thus $\partial^+(\alpha + \beta) \subseteq \partial^+(\alpha) \cup \partial^+(\beta)$.
- In the same way, from equation (4), for all $w \in \Sigma^+$, the following holds:
$$\partial_w(\alpha \cap \beta) \subseteq \partial_w(\alpha) ⩀ \partial_w(\beta) \subseteq \partial^+(\alpha) ⩀ \partial^+(\beta),$$
and then $\partial^+(\alpha \cap \beta) \subseteq \partial^+(\alpha) ⩀ \partial^+(\beta)$.
- From equation (5), for all $w \in \Sigma^+$, the following holds:
$$\partial_w(\alpha\beta) \subseteq \partial_w(\alpha)\beta \cup \bigcup_{v \in \mathrm{suff}(w)} \partial_v(\beta) \subseteq \partial^+(\alpha)\beta \cup \partial^+(\beta),$$
and thus $\partial^+(\alpha\beta) \subseteq \partial^+(\alpha)\beta \cup \partial^+(\beta)$.
- Finally, from equation (6), for all $w \in \Sigma^+$, the following holds:
$$\partial_w(\alpha^\star) \subseteq \bigcup_{v \in \mathrm{suff}(w)} \partial_v(\alpha)\alpha^\star \subseteq \partial^+(\alpha)\alpha^\star,$$
therefore, we have that $\partial^+(\alpha) \subseteq \partial^+(\alpha)\alpha^\star$.

□

The next example shows that the inclusion $\partial^+(\alpha \cap \beta) \subseteq \partial^+(\alpha) ⩀ \partial^+(\beta)$ is strict in some cases.

Example 12. Consider again $\alpha_1 = (b + ab + aab + abab) \cap (ab)^\star$. We have $\partial^+(\alpha_1) = \{bab \cap b(ab)^\star,\ ab \cap b(ab)^\star,\ b \cap b(ab)^\star,\ ab \cap (ab)^\star,\ \varepsilon \cap (ab)^\star\}$. It is easy to see that $\partial^+((ab)^\star) = \{b(ab)^\star, (ab)^\star\}$. Now, with $\beta = (b + ab + aab + abab)$ and $\partial^+(\beta) = \{\varepsilon, b, ab, bab\}$, one has

$$\begin{aligned}\partial^+(\beta) ⩀ \partial^+((ab)^\star) = \{&bab \cap b(ab)^\star, ab \cap b(ab)^\star,\ b \cap b(ab)^\star,\\ &\varepsilon \cap b(ab)^\star,\ bab \cap (ab)^\star,\ ab \cap (ab)^\star,\ b \cap (ab)^\star,\ \varepsilon \cap (ab)^\star\}.\end{aligned}$$

Thus $\partial^+(\alpha_1) \neq \partial^+(b + ab + aab + abab) ⩀ \partial^+((ab)^\star)$.

The following proposition relates the function ∂^+ with the support π.

Proposition 13. *Given $\alpha \in \mathsf{RE}_\cap$, $\partial^+(\alpha) \subseteq \pi(\alpha)$.*

Proof. The proof proceeds by induction on the structure of α. It is trivial that $\partial^+(\emptyset) = \pi(\emptyset)$, $\partial^+(\varepsilon) = \pi(\varepsilon)$ and $\partial^+(a) = \pi(a)$, for a symbol $a \in \Sigma$. Assume that $\partial^+(\alpha) \subseteq \pi(\alpha)$ and $\partial^+(\beta) \subseteq \pi(\beta)$ holds, for $\alpha, \beta \in \mathsf{RE}_\cap$. For $\alpha + \beta$, one has

$$\partial^+(\alpha + \beta) \subseteq \partial^+(\alpha) \cup \partial^+(\beta) \subseteq \pi(\alpha) \cup \pi(\beta).$$

For $\alpha \cap \beta$, one has

$$\partial^+(\alpha \cap \beta) \subseteq \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta) \subseteq \pi(\alpha) \mathbin{\dot\cap} \pi(\beta).$$

For $\alpha\beta$, one has

$$\partial^+(\alpha\beta) \subseteq \partial^+(\alpha)\beta \cup \partial^+(\beta) \subseteq \pi(\alpha)\beta \cup \pi(\beta).$$

Finally, for $\alpha^\star$,

$$\partial^+(\alpha^\star) \subseteq \partial^+(\alpha)\alpha^\star \subseteq \pi(\alpha)\alpha^\star.$$

□

Since, for every regular expression $\alpha \in \mathsf{RE}_\cap$, the set $\pi(\alpha)$ is finite, Proposition 13 also proves that the set $\partial^+(\alpha)$ is finite. For regular expressions without intersection it is known that π and ∂^+ coincide [14]. Examples 3 and 12 show that there exists $\alpha \in \mathsf{RE}_\cap$ such that $\pi(\alpha) \neq \partial^+(\alpha)$. The following lemmas establish some conditions for the equality of $\pi(\alpha \cap \beta)$ and $\partial^+(\alpha \cap \beta)$ to hold for $\alpha, \beta \in \mathsf{RE}_\cap$, and they will be used in Proposition 16.

Lemma 14. *Given $\alpha, \beta \in \mathsf{RE}_\cap$, one has $\pi(\alpha \cap \beta) = \partial^+(\alpha \cap \beta)$ if and only if $\pi(\alpha) = \partial^+(\alpha)$, $\pi(\beta) = \partial^+(\beta)$ and $\partial^+(\alpha \cap \beta) = \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta)$.*

Proof.
($\Rightarrow$) We have that $\pi(\alpha \cap \beta) = \partial^+(\alpha \cap \beta) \subseteq \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta)$. From Proposition 13 it follows that $\partial^+(\alpha) \subseteq \pi(\alpha)$ and $\partial^+(\beta) \subseteq \pi(\beta)$. Suppose, by contradiction, that $\partial^+(\alpha) \subsetneq \pi(\alpha)$ or $\partial^+(\beta) \subsetneq \pi(\beta)$. Then

$$\partial^+(\alpha \cap \beta) \subseteq \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta) \subsetneq \pi(\alpha) \mathbin{\dot\cap} \pi(\beta) = \pi(\alpha \cap \beta),$$

a contradiction since $\pi(\alpha \cap \beta) = \partial^+(\alpha \cap \beta)$. Thus, we conclude that $\pi(\alpha) = \partial^+(\alpha)$ and $\pi(\beta) = \partial^+(\beta)$. Consequently, $\pi(\alpha \cap \beta) = \pi(\alpha) \mathbin{\dot\cap} \pi(\beta) = \partial^+(\alpha \cap \beta)$.

($\Leftarrow$) This follows trivially from the definition of support, i.e., $\pi(\alpha \cap \beta) = \pi(\alpha) \mathbin{\dot\cap} \pi(\beta)$, since $\pi(\alpha) = \partial^+(\alpha)$ and $\pi(\beta) = \partial^+(\beta)$. □

Lemma 15. *Given $\alpha, \beta \in \mathsf{RE}_\cap$ such that $\partial_w(\alpha) = \pi(\alpha)$ or $\partial_w(\beta) = \pi(\beta)$ holds for all $w \in \Sigma^+$, then $\partial^+(\alpha \cap \beta) = \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta)$.*

Proof. First, note that if $\gamma \in \mathsf{RE}_\cap$ and $\partial_w(\gamma) = \pi(\gamma)$ for every $w \in \Sigma^+$, then $\partial^+(\gamma) = \bigcup_{w \in \Sigma^+} \partial_w(\gamma) = \pi(\gamma)$. It is enough to assume $\partial_w(\alpha) = \pi(\alpha)$, for all $w \in \Sigma^+$, in which case one has

$$\begin{aligned}
\partial^+(\alpha \cap \beta) &= \bigcup_{w\in\Sigma^+} (\partial_w(\alpha) \mathbin{\dot\cap} \partial_w(\beta) = \bigcup_{w\in\Sigma^+} (\pi(\alpha) \mathbin{\dot\cap} \partial_w(\beta)) \\
&= \bigcup_{w\in\Sigma^+} \{\ \alpha_i \cap \beta_j \mid \alpha_i \in \pi(\alpha),\ \beta_j \in \partial_w(\beta)\ \} \\
&= \left\{ \alpha_i \cap \beta_j \;\middle|\; \alpha_i \in \pi(\alpha),\ \beta_j \in \bigcup_{w\in\Sigma^+} \partial_w(\beta) \right\} \\
&= \{\ \alpha_i \cap \beta_j \mid \alpha_i \in \pi(\alpha),\ \beta_j \in \partial^+(\beta)\ \} \\
&= \pi(\alpha) \mathbin{\dot\cap} \partial^+(\beta) = \partial^+(\alpha) \mathbin{\dot\cap} \partial^+(\beta).
\end{aligned}$$

□

By Proposition 13, $|\pi(\alpha)|$ is an upper bound for the cardinality of $\partial^+(\alpha)$. This upper bound can be reached, as shown by the following proposition.

Proposition 16. *For any $n \in \mathbb{N}$ there exists a regular expression $r_n \in \mathsf{RE}_\cap$ of size $O(n)$ such that $|\partial^+(r_n)| = 2^{|r_n|_\Sigma - |r_n|_\cap - 1}$.*

Proof. Consider the regular expressions $r_n \in \mathsf{RE}_\cap$ from Example 7. We prove that $\pi(r_n) = \partial^+(r_n)$. The proof proceeds by induction on n. For $n = 0$ and for all $w \in \Sigma^+$, we have $\partial_w(a^\star a) = \{a^\star a, \epsilon\} = \partial^+(a^\star a) = \pi(a^\star a)$. Let us assume that $\pi(r_n) = \partial^+(r_n)$, for $n \geq 1$. It follows from Lemma 15 that

$$\partial^+(r_{n+1}) = \partial^+(r_n \cap a^\star a) = \partial^+(r_n) \mathbin{\dot\cap} \partial^+(a^\star a).$$

Since $\pi(a^\star a) = \partial^+(a^\star a)$, $\pi(r_n) = \partial^+(r_n)$, and $\partial^+(r_n \cap a^\star a) = \partial^+(r_n) \mathbin{\dot\cap} \partial^+(r_n)$, we conclude, from Lemma 14, that $\pi(r_{n+1}) = \pi(r_n \cap a^\star a) = \partial^+(r_n \cap a^\star a) = \partial^+(r_{n+1})$. □

The next example provides a non-trivial family of regular expressions for which the set of partial derivatives coincides with the support. Note that although their size grows exponentially the upper bound is not reached in this case.

Example 17. For $n \geq 0$ let the regular expression $s_n \in \mathsf{RE}_\cap$ be inductively defined by $s_0 = (a+b)^\star b(a+b)^\star$ and, for $n \geq 1$,

$$s_n = ((a+b)s_{n-1}(a+b)) \cap ((a+b)^\star(a+b)).$$

It is easy to see that $|s_n|_\Sigma = 5+8n$ and $|s_n|_\cap = n$. One has, $|\pi(s_0)| = |\{s_0, (a+b)^\star\}| = 2$ and $|\pi(s_1)| = |\{s_0(a+b), (a+b)^\star, \varepsilon\} \mathbin{\dot\cap} \{(a+b)^\star(a+b), \varepsilon\}| = 6$. For $n \geq 2$,

$$\begin{aligned}
|\pi(s_n)| &= |\,(\{s_{n-1}(a+b), \varepsilon\} \cup \pi(s_{n-1})(a+b)) \mathbin{\dot\cap} \{(a+b)^\star(a+b), \varepsilon\}| \\
&= 2(2 + |\pi(s_{n-1})|) \\
&= \sum_{i=2}^{n} 2^i + 3 \cdot 2^n.
\end{aligned}$$

The second equality holds because, for $n \geq 2$, $s_{n-1} \notin \pi(s_{n-1})$. Thus, we have $|\pi(s_n)| = O(2^n)$ for $n \geq 2$. Let $m = |s_n|_\Sigma - |s_n|_\cap - 1 = 5 + 7n - 1$, i.e. $n = (m-4)/7$.

Then, $|\pi(s_n)| = O(2^{\frac{1}{7}m}) = O(1.105^m)$, which is much smaller than the upper bound 2^m. However, $\pi(s_n) = \partial^+(s_n)$, for all $n \geq 0$, as we now prove by induction on n. The cases $n = 0$ and $n = 1$ are obvious, because the equality holds for regular expressions without intersections. Let us assume that $\pi(s_n) = \partial^+(s_n)$, for some $n \geq 2$. Let $s_{n+1} = r_{n+1} \cap t_0$, where $r_{n+1} = (a+b)s_n(a+b)$ and $t_0 = (a+b)^\star(a+b)$. It is clear that $\partial^+(t_0) = \pi(t_0)$. One has

$$\begin{aligned}\pi(r_{n+1}) &= \pi(a+b)s_n(a+b) \cup \pi(s_n(a+b))\\ &= \{s_n(a+b)\} \cup \pi(s_n(a+b)).\end{aligned}$$

Since $\pi(s_n) = \partial^+(s_n)$ and $\pi(a+b) = \partial^+(a+b)$, from Lemma 14 we have $\pi(s_n(a+b)) = \partial^+(s_n(a+b))$. Given that $\partial_a(r_{n+1}) = \partial_b(r_{n+1}) = \{s_n(a+b)\}$, then

$$\begin{aligned}\partial^+(r_{n+1}) &= \partial_a(r_{n+1}) \cup \partial_b(r_{n+1}) \cup \bigcup_{w\in\Sigma^+\setminus\{a,b\}} \partial_w(r_{n+1})\\ &= \{s_n(a+b)\} \cup \partial^+(s_n(a+b))\\ &= \pi(r_{n+1}).\end{aligned}$$

From Lemma 15, we have $\partial^+(r_{n+1} \cap t_0) = \partial^+(r_{n+1}) \mathbin{\dot{\cap}} \partial^+(t_0)$ and from Lemma 14 we conclude that $\pi(r_{n+1} \cap t_0) = \partial^+(r_{n+1} \cap t_0)$, i.e., $\pi(s_{n+1}) = \partial^+(s_{n+1})$.

5. Average Complexity Results

We know that the number of states in the partial derivative automaton of an expression α has $|\pi(\alpha)|$ as its tight upper bound. In this section we estimate an upper bound for the asymptotic average size of $\pi(\alpha)$. This is done using standard methods of analytic combinatorics as expounded by Flajolet and Sedgewick [17], that apply to generating functions $f(z) = \sum_n a_n z^n$ associated with combinatorial classes. Given some measure of the objects of a combinatorial class $\mathcal{A}$, the coefficient a_n represents the sum of the values of this measure for all objects of size n. We will use the notation $[z^n]f(z)$ for a_n. For an introduction to this approach applied to formal languages, we refer to Broda *et al.* [8].

Although the methods here used are the standard ones from the Analytic Combinatorics (and Complex Analysis), each application of these techniques is always a challenge, as one cannot foresee the analytic difficulties that one can incur into when conducting the study of the generation function.

The generating function f can be seen as a complex function, analytic in a neighbourhood of the origin, and the study of its behaviour near its dominant singularity ρ (in case there is only one, as it happens with the functions here considered) gives us access to the asymptotic form of its coefficients. In particular, if $f(z)$ is analytic in some appropriate neighbourhood of 0 containing ρ, then one has the following [17, 8]:

Proposition 18. *If* $f(z) = a - b\sqrt{1 - z/\rho} + o\left(\sqrt{1 - z/\rho}\right)$, *with* $a, b \in \mathbb{R}$, $b \neq 0$, *then*

$$[z^n]f(z) \sim \frac{b}{2\sqrt{\pi}}\,\rho^{-n} n^{-3/2}.$$

If $f(z) = \frac{a}{\sqrt{1-z/\rho}} + o\left(\frac{1}{\sqrt{1-z/\rho}}\right)$*, with* $a \in \mathbb{R}$*, and* $a \neq 0$*, then*

$$[z^n]f(z) \sim \frac{a}{\sqrt{\pi}}\,\rho^{-n}n^{-1/2}.$$

5.1. Counting Expressions and Symbols in General Regular Expressions

The average number of occurrences of particular symbols in standard regular expressions was already studied by Nicaud [25], as well as for special regular expressions by Broda et al. [9]. We present here the result of applying the methods used in those works to general regular expression with any number of unary and binary operators. Let us suppose that we have $s = s_n + s_c$ characters (letters or constants), $u = u_n + u_c$ unary operators and $b = b_n + b_c$ binary operators, where the index c denotes the number of objects that one is interested in counting, whereas the index n denotes the remaining ones. For example, in order to count the number of $+$ symbols occurring in regular expressions with an alphabet of k letters, these parameters would be $s_c = 0$ (no characters are to be counted), $s_n = k+1$ (because ε can also occur in the regular expressions along with the k different letters), $u_c = 0$, $u_n = 1$ (the $\star$ operator), $b_n = 1$ (the concatenation operator) and $b_c = 1$ (the $+$ operator).

The generating function $R(z)$ for the number of regular expressions of a given size satisfies the following relation

$$R(z) = sz + uzR(z) + bzR^2(z),$$

which yields

$$R(z) = \frac{1 - uz - \sqrt{\Delta(z)}}{2bz},$$

where $\Delta(z) = (u^2 - 4bs)z^2 - 2uz + 1$.

Analogously, the bivariate generating function $S(v,z)$, whose coefficients give the number of objects that simultaneously have a certain size and a fixed number of the things one wishes to count, satisfies

$$S(v,z) = (s_n + s_c v)z + (u_n + u_c v)zS(v,z) + (b_n + b_c v)zS^2(v,z),$$

from which one can derive the cumulative generating function $S(z)$, whose coefficients are the number of objects one wants to count occurring in the expressions of a certain size:

$$S(z) = \frac{\phi(z) + \psi(z)\sqrt{\Delta(z)}}{2b^2z\sqrt{\Delta(z)}},$$

where

$$\phi(z) = \left(2s_cb^2 - 2sbb_c - uu_cb + u^2b_c\right)z^2 + (u_cb - 2ub_c)z + b_c$$
$$\psi(z) = (ub_c - u_cb)z - b_c.$$

The smallest positive root of $\Delta(z)$ is the main singularity of both $R(z)$ and $S(z)$, and is given by

$$\rho = \rho_{s,u,b} = \begin{cases} \frac{1}{u+2\sqrt{bs}}, & \text{when } 4bs \neq u^2, \\ \frac{1}{2u}, & \text{when } 4bs = u^2. \end{cases}$$

Now, using the methods of analytic combinatorics as expounded in Broda et al. [8], for example, one obtains

$$[z^n]R(z) \sim \frac{\sqrt{2-2u\rho}}{4b\rho\sqrt{\pi}}\,\rho^{-n}n^{-\frac{3}{2}},$$

$$[z^n]S(z) \sim \frac{\phi(\rho)}{2b^2\rho\sqrt{\pi}\sqrt{2-2u\rho}}\,\rho^{-n}n^{-\frac{1}{2}}.$$

5.2. Counting Expressions, Letters and ∩ Symbols

The study of the combinatorial behaviour of the $\mathsf{RE}_\cap$-expressions, both in terms of the number of expressions and the number of occurring letters, can now be done using the results of the previous section. In this case, the parameters for the above mentioned results should be: $s_c = k, s_n = 1, u_c = 0, u_n = 1, b_c = 0$ and $b_n = 3$, giving rise to

$$[z^n]R_k(z) \sim \frac{\sqrt{2-2\rho_k}}{12\rho_k\sqrt{\pi}}\,\rho_k^{-n}n^{-\frac{3}{2}}, \tag{7}$$

and

$$[z^n]L_k(z) \sim \frac{k\rho_k}{\sqrt{2-2\rho_k}\;\sqrt{\pi}}\rho_k^{-n}n^{-\frac{1}{2}}, \tag{8}$$

where $\rho_k = \frac{1}{1+2\sqrt{3k+3}}$.

The average number of letters in an expression of size n is given by $\frac{[z^n]L_k(z)}{[z^n]R_k(z)}$. Using equations (7) and (8), one obtains, asymptotically,

$$|\alpha|_\Sigma \sim \frac{6k\rho_k^2}{1-\rho_k}\,|\alpha| \xrightarrow[k\to\infty]{} \frac{1}{2}|\alpha|. \tag{9}$$

The number of intersections in the $\mathsf{RE}_\cap$-expressions under consideration can be computed, again using the previous result, now with $s_c = 0, s_n = k+1, u_c = 0, u_n = 1, b_c = 1$ and $b_n = 2$, yielding

$$[z^n]I_k(z) \sim \frac{(k+1)\rho_k}{3\sqrt{\pi}\;\sqrt{2-2\rho_k}}\,\rho_k^{-n}n^{-\frac{1}{2}}. \tag{10}$$

The average number of symbols ∩ in an expression of size n is given by $\frac{[z^n]I_k(z)}{[z^n]R_k(z)}$. Using equations (7) and (10), one obtains, asymptotically,

$$|\alpha|_\cap \sim \frac{2(k+1)\rho_k^2}{1-\rho_k}\,|\alpha| \xrightarrow[k\to\infty]{} \frac{1}{6}|\alpha|. \tag{11}$$

5.3. Average Size of π

Let $P_k(z)$ denote the generating function for the size of $\pi(\alpha)$ for expressions without $\emptyset$. From Definition 2 it follows that, given an expression α, an upper bound[3] $p(\alpha)$ for the number of elements in the set $\pi(\alpha)$ satisfies:

$$\begin{aligned} p(\varepsilon) &= 0, & p(\alpha+\beta) &= p(\alpha)+p(\beta), \\ p(a) &= 1, \quad \text{for } a \in \Sigma, & p(\alpha\beta) &= p(\alpha)+p(\beta), \\ p(\alpha^\star) &= p(\alpha), & p(\alpha\cap\beta) &= p(\alpha)p(\beta). \end{aligned} \tag{12}$$

From this, we get

$$\begin{aligned} P_k(z) = \sum_\alpha p(\alpha)z^{\|\alpha\|} &= \sum_a p(a)z + \sum_\alpha p(\alpha^\star)z^{\|\alpha\|+1} \\ &+ \sum_{\alpha,\beta} p(\alpha+\beta)z^{\|\alpha\|+\|\beta\|+1} + \sum_{\alpha,\beta} p(\alpha\beta)z^{\|\alpha\|+\|\beta\|+1} \\ &+ \sum_{\alpha,\beta} p(\alpha\cap\beta)z^{\|\alpha\|+\|\beta\|+1}. \end{aligned}$$

Noticing that

$$\begin{aligned} \sum_{\alpha,\beta} p(\alpha+\beta)z^{\|\alpha+\beta\|} &= \sum_{\alpha,\beta}(p(\alpha)+p(\beta))z\,z^{\|\alpha\|}z^{\|\beta\|} = \\ &= z\left(\sum_{\alpha,\beta} p(\alpha)z^{\|\alpha\|}z^{\|\beta\|} + \sum_{\alpha,\beta} p(\beta)z^{\|\beta\|}z^{\|\alpha\|}\right) \\ &= 2z\left(\sum_\alpha p(\alpha)z^{\|\alpha\|}\right)\left(\sum_\beta z^{\|\beta\|}\right) \; = \; 2P_k(z)R_k(z), \end{aligned}$$

it is, then, easy to conclude that $P_k(z) = kz + 4zP_k(z)R_k(z) + zP_k(z) + zP_k(z)^2$, from which we obtain the following closed expression

$$P_k(z) = \frac{1 - z + 2\sqrt{q_k(z)} - \sqrt{p_k(z) + 4(1-z)\sqrt{q_k(z)}}}{6z}, \tag{13}$$

where

$$p_k(z) = 5 - 10z - (43+84k)z^2, \text{ and } q_k(z) = 1 - 2z - (11+12k)z^2.$$

The roots of $q_k(z)$ are

$$\rho_k = \frac{1}{1+2\sqrt{3+3k}}, \text{ and } \bar\rho_k = \frac{1}{1-2\sqrt{3+3k}}.$$

One now needs to determine the dominant singularity of $P_k(z)$, which can either be the positive root of $q_k(z)$ or a root of $r_k(z) = p_k(z) + 4(1-z)\sqrt{q_k(z)}$. Thus, we need

[3] This upper bound corresponds to the case where all unions in $\pi(\alpha)$ are disjoint.

to know which of the two expressions $r_k(z)$ or $q_k(z)$ has the smallest positive zero. Because this is not trivial (note that one needs to decide this for all k), one will do it indirectly using the method expounded in the following paragraphs.

Observing that $r_k(0) = 9$ is positive, and

$$r_k(\rho_k) = \frac{12\left(13 - 14k - 24k^2 + (8k-4)\sqrt{3+3k}\right)}{(11+12k)^2} < 0,$$

by Bolzano theorem, $r_k(z)$ must have a positive zero smaller than ρ_k. Noticing that

$$r_k(\bar{\rho}_k) = -\frac{12\left(-13 + 14k + 24k^2 + (8k-4)\sqrt{3+3k}\right)}{(11+12k)^2} < 0,$$

one concludes that $r_k(z)$ has necessarily two real zeros in its domain, $[\bar{\rho}_k, \rho_k]$. Analogously, $s_k(z) = p_k(z) - 4(1-z)\sqrt{q_k(z)}$ has also two real zeros in the same interval, and since $r_k(z)s_k(z)$ is a fourth degree polynomial, it follows that $r_k(z)$ has exactly two zeros, η_k and η'_k, which are real. Since $s_k(0) = 1 < r_k(0) = 9$, and $r_k(x) = s_k(x)$ only at the end points of $[\bar{\rho}_k, \rho_k]$ it follows that $s_k(x) < r_k(x)$ in $]\bar{\rho}_k, \rho_k[$. Considering the four real zeros of the polynomial $r_k(z)s_k(z)$, given what we just said, we conclude that the two more distant zeros from the origin are the roots of $r_k(z)$. In fact, we can obtain an explicit expression for the zeros of $r_k(z)s_k(z)$ by noticing that

$$\begin{aligned} p_k(z) \pm 4(1-z)\sqrt{q_k(z)} &= \left(1 - z \pm 2\sqrt{q_k(z)}\right)^2 - 36kz^2 \\ &= \left(1 - z \pm 2\sqrt{q_k(z)} - 6\sqrt{k}z\right)\left(1 - z \pm 2\sqrt{q_k(z)} + 6\sqrt{k}z\right), \end{aligned}$$

and thus, solving the equations resulting of nulling those factors, we obtain the four zeros of $r_k(z)s_k(z)$:

$$\begin{aligned} \eta_k &= \frac{4\sqrt{2k+1} + 2\sqrt{k} - 1}{28k + 4\sqrt{k} + 15}, & \eta'_k &= -\frac{4\sqrt{2k+1} + 2\sqrt{k} + 1}{28k - 4\sqrt{k} + 15}, \\ \eta''_k &= \frac{4\sqrt{2k+1} - 2\sqrt{k} - 1}{28k - 4\sqrt{k} + 15}, & \eta'''_k &= -\frac{4\sqrt{2k+1} - 2\sqrt{k} + 1}{28k + 4\sqrt{k} + 15}. \end{aligned} \tag{14}$$

It is possible to verify that η_k and η'_k are the roots of $r_k(z)$ and the other two the roots from $s_k(z)$. Therefore, one has

$$r_k(z)s_k(z) = (7056k^2 + 7416k + 2025)(z - \eta_k)(z - \eta'_k)(z - \eta''_k)(z - \eta'''_k). \tag{15}$$

From (13) one has

$$6zP_k(z) = 1 - z - \sqrt{r_k(z)} + 2\sqrt{q_k(z)}, \tag{16}$$

and we split the study of the coefficients of the series of $P_k(z)$ into the study of the coefficients of $1 - z - \sqrt{r_k(z)}$ and of $2\sqrt{q_k(z)}$. For the first one, we use that

$$r_k(z) = \frac{7056k^2 + 7416k + 2025}{s_k(z)}\eta_k(\eta'_k - z)(\eta''_k - z)(\eta'''_k - z)\left(1 - \frac{z}{\eta_k}\right),$$

and the fact that, given a complex function f defined on a neighbourhood of η such that $\lim_{z\to\eta} f(z) = a$, one has, for all $r \in \mathbb{R}$, $f(z)(1-z/\eta)^r = a(1-z/\eta)^r + o((1-z/\eta)^r)$, together with Proposition 18, to obtain

$$[z^n]\left(1 - z - \sqrt{r_k(z)}\right) \sim \lambda_k \eta_k^{-n} n^{-\frac{3}{2}},$$

where

$$\lambda_k = \left(\frac{(7056k^2 + 7416k + 2025)(\eta'_k - \eta_k)(\eta''_k - \eta_k)(\eta'''_k - \eta_k)\eta_k}{2\pi s_k(\eta_k)}\right)^{\frac{1}{2}}. \tag{17}$$

For the last summand one has, similarly,

$$2\sqrt{q_k(z)} = 4\sqrt[4]{3+3k}\,\rho_k^{\frac{1}{2}}(\rho_k - \bar{\rho}_k)^{\frac{1}{2}}(1 - z/\rho_k)^{\frac{1}{2}} + o\left((1 - z/\rho_k)^{\frac{1}{2}}\right),$$

from which it follows, $[z^n]2\sqrt{q_k(z)} \sim -\mu_k \rho_k^{-n} n^{-\frac{3}{2}}$, where

$$\mu_k = 2\pi^{-\frac{1}{2}} \rho_k^{\frac{1}{2}} \sqrt[4]{3+3k}. \tag{18}$$

Summing up, we get that

$$[z^n]P_k(z) \sim \frac{1}{6}\left(\lambda_k \eta_k^{-(n+1)} - \mu_k \rho_k^{-(n+1)}\right) n^{-\frac{3}{2}}. \tag{19}$$

In order to see what this result entails for the average case when compared with the worst case result, expressed in Proposition 5, attend to the following

$$\left(\frac{[z^n]P_k(z)}{[z^n]R_k(z)}\right)^{\frac{1}{n}} \sim \left(\frac{\frac{1}{6}\lambda_k \eta_k^{-(n+1)} n^{-\frac{3}{2}}}{c_k \rho_k^{-n-\frac{1}{2}}(n+1)^{-\frac{3}{2}}}\right)^{\frac{1}{n}} \xrightarrow[n\to\infty]{} \frac{\rho_k}{\eta_k}.$$

Setting $\gamma_k = \frac{\rho_k}{\eta_k}$, this means that, on average,

$$|\pi(\alpha)| \sim \gamma_k^{|\alpha|}.$$

One has $\gamma_1 \sim 1.00495$, $\gamma_2 \sim 1.01655$, $\gamma_{10} \sim 1.04137$, $\gamma_{100} \sim 1.05294$, and

$$\lim_{k\to\infty} \gamma_k = \frac{7\sqrt{3}}{6\sqrt{2}+3} \sim 1.05564.$$

Proposition 19. *For large values of k and n, an upper bound for the average number of states of $\mathcal{A}_{pd}$ is $(1.056 + o(1))^n$.*

Considering the estimates given in (9) and (11), the worst-case upper bound $2^{|\alpha|_\Sigma - |\alpha|_\cap - 1}$ from Proposition 5 leads to an upper bound for the average case roughly of $\sqrt[3]{2}^{|\alpha|}$, for α large enough. As $\sqrt[3]{2} \sim 1.25992$, the result just obtained shows that the upper bound for the average complexity is significantly smaller than the one for the worst case.

Table 1: Experimental Results.

k	$\|\alpha\|$	$\|\alpha\|_\Sigma$	$\|\alpha\|_\cap$	$\emptyset$	$\|\delta_{\mathsf{pd}}\|$	$\|\partial(\alpha)\|$	$\|\pi(\alpha)\|$
1	25	5.42 12	3.26 10	0.17 1	2.86 43	2.50 14	1.80 13
	50	10.59 20	6.73 17	0.18 1	4.52 420	3.06 27	2.70 40
	100	20.99 34	13.69 25	0.19 1	7.13 744	3.70 51	4.18 175
	150	31.39 50	20.54 36	0.19 1	9.93 3691	4.28 115	5.93 324
	200	41.79 59	27.47 44	0.20 1	11.40 6330	4.49 288	7.46 621
	300	62.65 88	41.34 64	0.20 1	15.38 3660	5.32 1873	11.98 9360
2	25	7.46 13	3.41 9	0.28 1	2.75 42	2.47 14	2.94 24
	50	14.60 23	6.96 16	0.31 1	3.52 76	2.77 14	5.32 80
	100	28.88 42	14.14 26	0.29 1	4.37 573	3.07 39	11.82 566
	150	43.12 58	21.19 37	0.31 1	4.77 397	3.16 63	21.81 3949
	200	57.44 75	28.41 46	0.31 1	4.74 214	3.22 41	40.03 9250
	300	86.00 106	42.73 64	0.32 1	5.08 493	3.28 57	121.78 134604
5	25	9.73 13	3.56 10	0.40 1	2.58 40	2.42 11	4.65 40
	50	19.04 25	7.28 17	0.42 1	2.82 88	2.52 16	10.59 193
	100	37.66 46	14.72 29	0.42 1	3.14 136	2.65 20	36.85 3650
	150	56.34 69	22.16 37	0.42 1	3.34 96	2.71 21	116.20 21216
	200	74.96 89	29.64 47	0.43 1	3.29 119	2.68 22	356.54 72135
	300	112.19 129	44.61 65	0.43 1	3.43 152	2.72 22	3470.22 2506175
10	25	10.90 13	3.63 10	0.46 1	2.37 41	2.38 11	5.72 30
	50	21.33 25	7.46 16	0.48 1	2.63 49	2.47 16	14.72 168
	100	42.26 50	15.16 28	0.50 1	2.73 98	2.50 22	68.91 10104
	150	63.14 72	22.87 39	0.48 1	2.88 72	2.58 23	302.82 80080
	200	84.13 95	30.43 49	0.49 1	2.88 115	2.58 22	1269.14 392148
	300	125 138	45.68 67	0.47 1	2.91 59	2.58 19	18369 5637926

5.4. Experimental Results

In order to compare the size of the partial derivative automaton and the size of the support of the corresponding regular expression, we ran some experiments, using the FAdo package [16]. For the results to be statistically significant, regular expressions were uniformly random generated using a version of the grammar for $\mathsf{RE}_\cap$ in prefix notation. For each size $n \in \{25, 50, 100, 150, 200, 300\}$ and alphabet size $k \in \{1, 2, 5, 10\}$ samples of 10000 regular expressions were generated[4]. This is sufficient to ensure a 95% confidence level within a 1% error margin.

For each sample we computed the average and the maximum value of several measures, which are presented in Table 1. For each n and k, the first row has the average values and the second the maximal values. The column labelled with $\emptyset$ indicates the ratio of expressions which are equivalent to the empty language. The column labelled with $|\delta_{\mathsf{pd}}|$ indicates the number of transitions of the partial derivative automaton. As it is evident, the set of partial derivatives is on average much smaller than the support. We note that $\partial(\alpha)$ includes α, while $\pi(\alpha)$ may not include it. Even for small alphabets it seems that the size of $\mathcal{A}_{pd}$ does not grow exponentially.

6. A Special Case

We now consider a restricted form of regular expressions with intersection, and study the complexity of the conversion to equivalent partial derivative automata. Let $\mathsf{RE}_{!\cap} \subseteq \mathsf{RE}_{\cap}$ be the set of regular expressions that have exactly one intersection at the top level (or is $\emptyset$). This set is generated by the following grammar:

$$\begin{aligned}&\alpha \to (\beta \cap \beta) \mid \emptyset\\ &\beta \to \varepsilon \mid a \mid (\beta + \beta) \mid (\beta \cdot \beta) \mid (\beta)^* \qquad (a \in \Sigma).\end{aligned}$$

Note that β corresponds to a standard regular expression with the exclusion of $\emptyset$, i.e $\beta \in \mathsf{RE} \setminus \{\emptyset\}$. For $\alpha \in \mathsf{RE}_{!\cap}$, with $\alpha = \beta_1 \cap \beta_2$ and $\beta_i \in \mathsf{RE} \setminus \{\emptyset\}$, the support of α is $\pi(\alpha) = \pi(\beta_1) \Cap \pi(\beta_2)$, where $\pi(\beta_i)$ can be computed using Definition 2 without the intersection case. We know that $|\pi(\beta)| \leq |\beta|_\Sigma$, $\beta \in \mathsf{RE}$. In which case, the asymptotic average size was studied in [6].

The next proposition gives an upper-bound for the support of $\alpha \in \mathsf{RE}_{!\cap}$.

Proposition 20. *For all $\alpha \in \mathsf{RE}_{!\cap}$, the inequality $|\pi(\alpha)| \leq \left(\frac{|\alpha|_\Sigma}{2}\right)^2$ holds.*

Proof. For $\alpha = \emptyset$ the result is trivially true. Otherwise, $\alpha = \beta_1 \cap \beta_2$ with $\beta_i \in \mathsf{RE} \setminus \{\emptyset\}$. In this case, we know that $|\pi(\beta_1)| \leq |\beta_1|_\Sigma$ and $|\pi(\beta_2)| \leq |\beta_2|_\Sigma$. Thus,

$$|\pi(\alpha)| = \pi(\beta_1) \Cap \pi(\beta_2) \leq |\pi(\beta_1)| \cdot |\pi(\beta_2)| \leq |\beta_1|_\Sigma \cdot |\beta_2|_\Sigma.$$

Since $|\alpha|_\Sigma = |\beta_1|_\Sigma + |\beta_2|_\Sigma$, the value of $|\beta_1|_\Sigma \cdot |\beta_2|_\Sigma$ is maximized when $|\beta_1|_\Sigma = |\beta_2|_\Sigma = \frac{|\alpha|_\Sigma}{2}$. Consequently, $|\pi(\alpha)| \leq \left(\frac{|\alpha|_\Sigma}{2}\right)\left(\frac{|\alpha|_\Sigma}{2}\right) = \left(\frac{|\alpha|_\Sigma}{2}\right)^2$. □

The next example shows that this upper bound is reached.

Example 21. Consider

$$s_n = \underbrace{a^\star \cdots a^\star}_{n \geq 1} \in \mathsf{RE}.$$

[4] For $n = 300$ and $k = 10$ only 2000 regular expressions were used in the calculations due to the huge values involved.

Since $\pi(s_n) = \{s_n, s_{n-1}, \ldots, s_1\}$, we have $|\pi(s_n)| = |s_n|_\Sigma = n$.

Now, let $r_n \in \mathsf{RE}_{!\cap}$ be a regular expression defined as $r_n = s_n \cap s_n$, for $n \geq 1$. Then, the size of the support of r_n is $|\pi(r_n)| = |\pi(s_n) \mathbin{\bar{\cap}} \pi(s_n)| = |\pi(s_n)| \cdot |\pi(s_n)| = n^2$. Since, $|r_n|_\Sigma = 2n$, we have that $|\pi(r_n)| = n^2 = \left(\frac{|r_n|_\Sigma}{2}\right)^2$, which is exactly the upper bound given in Proposition 20.

Although the upper bound is reached, for this restricted set of regular expressions, examples 3 and 12 show that there are expressions α of this type for which $\partial^+(\alpha) \subsetneq \pi(\alpha)$.

6.1. Average Size of π

In this section we estimate an upper bound for the asymptotic average size of $\pi(\alpha)$, with $\alpha \in \mathsf{RE}_{!\cap}$, which provides an upper-bound for the average state complexity of $\mathcal{A}_{pd}(\alpha)$. The generating function for the size of $\beta \in \mathsf{RE}$ (excluding the $\emptyset$) over a k-ary alphabet is [26, 6]

$$S_k(z) = \frac{1 - z - \sqrt{\Delta_k(z)}}{4z},$$

where $\Delta_k(z) = 1 - 2z - (7 + 8k)z^2$, and the generating function for the number of letters in a standard expression is

$$L_k(z) = \frac{kz}{\sqrt{\Delta_k(z)}}.$$

It is easy to see that the generating function for the size of $\alpha \in \mathsf{RE}_{!\cap}$ is then

$$\begin{aligned} R'_k(z) &\sim zS_k^2(z) \\ &= \frac{(1 - z - \sqrt{\Delta_k(z)})^2}{16z} \\ &= \frac{1 - 2z - (4k+3)z^2 + (z-1)\sqrt{\Delta_k(z)}}{8z}. \end{aligned}$$

In the same way, the generating function for the number of letters in an expression $\alpha \in \mathsf{RE}_{!\cap}$ is $L'_k(z) = 2zL_k(z)S_k(z)$.

Using the same techniques as in Broda et al. [8, 10], one can obtain the following asymptotic estimates

$$[z^n]R'_k(z) \sim \frac{1}{2\sqrt{2\pi}}\sqrt{\rho_k(k+1)\sqrt{2k+2}}\ \rho_k^{-n} n^{-\frac{3}{2}}, \tag{20}$$

$$[z^n]L'_k(z) \sim \frac{k\sqrt{k+1}\ \rho_k^2}{\sqrt{\pi(1-\rho_k)}}\ \rho_k^{-n} n^{-\frac{1}{2}}, \tag{21}$$

where $\rho_k = \frac{1}{1+\sqrt{8+8k}}$.

From the above discussion, given $\alpha \in \mathsf{RE}_{!\cap}$, with $\alpha = \beta_1 \cap \beta_2$, an upper bound for the size of $\pi(\alpha)$ is $p'(\alpha) = p(\beta_1)p(\beta_2)$, where $p(\beta_i)$ is given by Equations (12) omitting the intersection case. Note that, in this case, when not considering intersections, p counts exactly the number of letters in a standard expression. Thus, the generating function for an upper bound for the size of $\pi(\alpha)$ is the rational function

$$P'_k(z) = zL_k(z)^2 = \frac{k^2z^3}{1-2z-(7+8k)z^2}.$$

To obtain an estimate for the average size of $|\pi(\alpha)|$ relative to $|\alpha|_\Sigma$, it only remains to compute $[z^n]P'_k(z)$. The roots of the denominator of $P'_k(z)$ are ρ_k and $\overline{\rho}_k = \frac{1}{1-\sqrt{8+8k}}$. It is easy to see that one has $(7+8k)\rho_k\overline{\rho}_k = -1$, and

$$P'_k(z) = \frac{k^2z^3}{(7+8k)(\overline{\rho}_k - z)(\rho_k - z)} = \frac{-k^2z^3}{(1-z/\rho_k)(1-z/\overline{\rho}_k)}.$$

Now,

$$\frac{1}{(1-z/\rho_k)(1-z/\overline{\rho}_k)} = \frac{A}{1-z/\rho_k} + \frac{B}{1-z/\overline{\rho}_k},$$

where $A = \frac{\overline{\rho}_k}{\overline{\rho}_k - \rho_k}$ and $B = \frac{\rho_k}{\rho_k - \overline{\rho}_k}$. Using all this, one easily gets

$$[z^n]P'_k(z) = \frac{k^2\rho_k\overline{\rho}_k}{\overline{\rho}_k - \rho_k}\left(\frac{1}{\rho_k^{n-2}} - \frac{1}{\overline{\rho}_k^{\,n-2}}\right).$$

From this one finally concludes that

$$\frac{[z^n]P'_k(z)}{[z^n]R'_k(z)} \sim \frac{\sqrt{2\pi}k^2\rho_k^{\frac{3}{2}}}{2\sqrt{k+1}(2k+2)^{\frac{3}{4}}}\, n^{\frac{3}{2}}.$$

Therefore, we have

$$\frac{[z^n]L'_k(z)}{[z^n]R'_k(z)} \sim \frac{2k\rho_k^2\sqrt{2k+2}}{\sqrt{\rho_k(1-\rho_k)(k+1)\sqrt{2k+2}}}\, n,$$

and

$$\frac{[z^n]P'_k(z)}{[z^n]L'_k(z)} \sim \frac{k\rho_k\overline{\rho}_k\sqrt{2\pi(1-\rho_k)}}{(\overline{\rho}_k - \rho_k)\sqrt{2k+2}}\, \sqrt{n}.$$

From this one concludes that the average size of $\pi(\alpha)$, for $\alpha \in \mathsf{RE}_{!\cap}$ of alphabetic size m, has an upper bound of the form $\xi_k m^{3/2}$, where

$$\xi_k = \frac{\sqrt{k\pi}\,(1-\rho_k)^{\frac{3}{4}}}{2^{\frac{25}{8}}\rho_k^{\frac{3}{4}}(k+1)^{\frac{7}{8}}} \to \frac{1}{8}\sqrt{2\pi}, \text{ as } k \to \infty.$$

Therefore, one has a significant improvement for the average case when compared with the worst case.

From all that was done above, the next result follows.

Proposition 22. *For large values of k and n an upper bound for the average number of states of $\mathcal{A}_{pd}(\alpha)$ for $\alpha \in \mathsf{RE}_{!\cap}$ is $O(|\alpha|_\Sigma^{\frac{3}{2}})$.*

7. Conclusions

The conversion of a regular expression with intersection, α, to an NFA is, in the worst-case, $2^{\Omega(|\alpha|)}$ [19, 22, 21]. This fact may lead one to believe that, although succinct, these expressions are not useful in practical applications. Here we show that, asymptotically, an upper bound for the average-state complexity of $\mathcal{A}_{pd}(\alpha)$ is exponential but with a base only slightly above 1. Some experimental results, using a uniform distribution, suggest that, on average, the upper bound may not be exponential. Considering regular expressions with only one intersection at the top level, the state complexity of $\mathcal{A}_{pd}(\alpha)$ is quadratic in the worst-case, and asymptotically, and on average, at most $O(n^{\frac{3}{2}})$. If we allow more than one intersection at the top level, Example 7 shows that the complexity of $\mathcal{A}_{pd}$ turns out to be exponential. The identification of other families of semi-extended regular expressions for which the conversion to NFAs is polynomial might be of practical interest.

References

[1] Valentin Antimirov. Partial derivatives of regular expressions and finite automaton constructions. *Theoret. Comput. Sci.*, 155(2):291–319, 1996.

[2] Valentin M. Antimirov and Peter D. Mosses. Rewriting extended regular expressions. In G. Rozenberg and A. Salomaa, editors, *1st DLT*, pages 195–209. World Scientific, 1994.

[3] Rafaela Bastos. Manipulation of extended regular expressions with derivatives. Master's thesis, Faculdade de Ciências da Universidade do Porto, 2015.

[4] Rafaela Bastos, Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. On the state complexity of partial derivative automata for regular expressions with intersection. In *18th DCFS*, volume 9777 of *LNCS*, pages 45–59. Springer, 2016.

[5] Rafaela Bastos, Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. On the average complexity of partial derivative automata for semi-extended expressions. *Journal of Automata, Languages and Combinatorics*, 22(1–3):5–28, 2017.

[6] Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. On the average state complexity of partial derivative automata. *Int. J. Found. Comput. S.*, 22(7):1593–1606, 2011.

[7] Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. On the average size of Glushkov and partial derivative automata. *Int. J. Found. Comput. S.*, 23(5):969–984, 2012.

[8] Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. A hitchhiker's guide to descriptional complexity through analytic combinatorics. *Theor. Comput. Sci.*, 528:85–100, 2014.

[9] Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. Partial derivative automaton for regular expressions with shuffle. In Jeffrey Shallit and Alexander Okhotin, editors, *17th DCFS*, volume 9118 of *LNCS*, pages 21–32. Springer, 2015.

[10] Sabine Broda, António Machiavelo, Nelma Moreira, and Rogério Reis. On the average complexity of strong star normal form. In Cezar Câmpeanu and Giovanni Pighizzini, ed-

itors, *19th erDescription Complexity of Formal Systems (DCFS 2017)*, LNCS. Springer, 2017.

[11] Janusz A. Brzozowski. Derivatives of regular expressions. *JACM*, 11(4):481–494, 1964.

[12] Pascal Caron, Jean-Marc Champarnaud, and Ludovic Mignot. Partial derivatives of an extended regular expression. In Adrian Horia Dediu, Shunsuke Inenaga, and Carlos Martín-Vide, editors, *5th LATA*, volume 6638 of *LNCS*, pages 179–191. Springer, 2011.

[13] Pascal Caron, Jean-Marc Champarnaud, and Ludovic Mignot. A general framework for the derivation of regular expressions. *RAIRO - Theor. Inf. and Applic.*, 48(3):281–305, 2014.

[14] Jean-Marc Champarnaud and Djelloul Ziadi. From Mirkin's prebases to Antimirov's word partial derivatives. *Fundam. Inform.*, 45(3):195–205, 2001.

[15] Tom Christiansen, brian d foy, Larry Wall, and Jon Orwant. *Programming Perl.* O'Reilly Media, 2012. 4th edition.

[16] Project FAdo. FAdo: tools for formal languages manipulation. `http://fado.dcc.fc.up.pt/`, Access date:1.11.2016.

[17] Philippe Flajolet and Robert Sedgewick. *Analytic Combinatorics.* CUP, 2008.

[18] Martin Fürer. The complexity of the inequivalence problem for regular expressions with intersection. In J. W. de Bakker and Jan van Leeuwen, editors, *7th ICALP*, volume 85 of *LNCS*, pages 234–245. Springer, 1980.

[19] Wouter Gelade. Succinctness of regular expressions with interleaving, intersection and counting. *Theor. Comput. Sci.*, 411(31-33):2987–2998, 2010.

[20] Wouter Gelade and Frank Neven. Succinctness of the complement and intersection of regular expressions. In Susanne Albers and Pascal Weil, editors, *25th STACS*, volume 1 of *LIPIcs*, pages 325–336. Schloss Dagstuhl — Leibniz-Zentrum für Informatik, Germany, 2008.

[21] Hermann Gruber. *On the descriptional and algorithmic complexity of regular languages.* PhD thesis, Justus Liebig University Giessen, 2010.

[22] Hermann Gruber and Markus Holzer. Finite automata, digraph connectivity, and regular expression size. In Luca Aceto, Ivan Damgård, Leslie Ann Goldberg, Magnús M. Halldórsson, Anna Ingólfsdóttir, and Igor Walukiewicz, editors, *35th ICALP*, volume 5126 of *LNCS*, pages 39–50. Springer, 2008.

[23] Tao Jiang and Bala Ravikumar. A note on the space complexity of some decision problems for finite automata. *Information Processing Letters*, 40(1):25–31, 1991.

[24] B. G. Mirkin. An algorithm for constructing a base in a language of regular expressions. *Engineering Cybernetics*, 5:51–57, 1966.

[25] Cyril Nicaud. *Étude du comportement en moyenne des automates finis et des langages rationnels.* PhD thesis, Université de Paris 7, 2000.

[26] Cyril Nicaud. On the average size of Glushkov's automata. In Adrian Horia Dediu, Armand-Mihai Ionescu, and Carlos Martín-Vide, editors, *3rd LATA*, volume 5457 of *LNCS*, pages 626–637. Springer, 2009.

[27] Holger Petersen. The membership problem for regular expressions with intersection is complete in LOGCFL. In Helmut Alt and Afonso Ferreira, editors, *19th STACS*, volume 2285 of *LNCS*, pages 513–522. Springer, 2002.

[28] Koushik Sen and Grigore Rosu. Generating optimal monitors for extended regular expressions. *Electr. Notes Theor. Comput. Sci.*, 2003.

[29] Eric van der Vlist. *RELAX NG.* O'Reilly Media, 2003.